MASTER'S THESIS

Simo Korpela

Comparing the Performance of Multivariate Location Tests for L_p-norm Distributed Data

UNIVERSITY OF TAMPERE
School of Information Sciences
Statistics

June 2014

University of Tampere

School of Information Sciences

KORPELA, SIMO: Comparing the Performance of Multivariate Location Tests for L_p-norm Distributed Data

Master's Thesis, 59 p., 11 appendix pages

Statistics

June 2014

Abstract

There are plenty of tests for multivariate location around which all make slightly different assumptions. The classical parametric procedure for the one-sample location problem is Hotelling's T^2 test (Hotelling, 1931) which assumes that the data is generated from a multivariate normal distribution. The test is optimal under the assumption of normality. However, the method may lead to unreliable results when the underlying distribution strongly deviates from the assumed model. As a reaction to this, the aim in research has been to develop methods that are valid under much weaker conditions than the normal-theory based Hotelling's T^2. Many nonparametric methods have been developed in the literature with the objective of extending to the multivariate context the classical univariate sign and rank techniques. Different attempts to generalize the classical nonparametric sign and rank methods have led to a huge body of literature.

This thesis presents the main multivariate tests of location and reports the results of an extensive simulation study. Tests based on marginal signs and ranks (Puri and Sen, 1971), spatial signs and ranks (Oja, 2010), Oja signs and ranks (Oja, 1999), the optimal signed-rank score tests by Hallin and Paindaveine (2002a, 2002b), and tests using marginal signs and ranks in the symmetric independent component (IC) model (Nordhausen, Oja, and Paindaveine, 2009) are discussed and applied. The parametric Hotelling's T^2 test serves as a reference test. The goal is to provide practical guidelines which test might be most useful in practice.

In our simulation study, the powers of the different location tests are compared under various settings, namely, under different underlying distributions, sample sizes, dimensions, and deviations from the null value. As extensions to normally distributed data, L_p-norm distributions are used as simulation data. We consider eleven different choices of underlying distributions, four different sample sizes, three different dimensions, and four different deviations from the null value. The proposed procedures are easy to implement on statistical programming languages such as R.

Keywords: multivariate location problem, multivariate distributions, nonparametric methods, affine invariance, affine equivariance

Contents

Abbreviations

$\sim$	distributed as
$'$	transpose
i.i.d.	independent and identically distributed
$\mathbf{I}_d$	$d \times d$ identity matrix
$\mathbf{O}$	orthogonal matrix
$\mathbf{J}$	sign-change matrix (a diagonal matrix with entries ± 1)
$\mathbf{P}$	permutation matrix (obtained by permuting the rows or columns of $\mathbf{I}_d$)
$\mathbf{\Sigma}$	scatter matrix
$\|\cdot\|$	vector norm of $\cdot$
$\|\cdot\|_p$	L_p-norm (or p-norm) of $\cdot$
$\Gamma(\cdot)$	gamma function
$U(d,p)$	L_p-norm uniform distribution
$S(d,p)$	L_p-norm spherical distribution
$N_d(\mathbf{0},\mathbf{I}_d,p)$	p-generalized normal distribution
ICS	invariant coordinate selection
IC model	independent component model
$\mathbf{S}_1$	marginal sign function
$\mathbf{R}_1$	marginal rank function
$\mathbf{Q}_1$	marginal signed-rank function
$\mathbf{S}_2$	spatial sign function
$\mathbf{R}_2$	spatial rank function
$\mathbf{Q}_2$	spatial signed-rank function
$\mathbf{S}_3$	Oja sign function
$\mathbf{R}_3$	Oja rank function
$\mathbf{Q}_3$	Oja signed-rank function
Q^2_{MS}	marginal sign test statistic
Q^2_{MR}	marginal Wilcoxon signed-rank test statistic
Q^2_{SS}	spatial sign test statistic
Q^2_{SR}	spatial signed-rank test statistic
Q^2_{OS}	Oja sign test statistic
Q^2_{OR}	Oja signed-rank test statistic
Q^2_{HPIS}	Hallin & Paindaveine sign test statistic based on interdirections
Q^2_{HPIR}	Hallin & Paindaveine signed-rank test statistic based on interdirections
Q^2_{HPTS}	Hallin & Paindaveine sign test statistic based on Tyler's angles
Q^2_{HPTR}	Hallin & Paindaveine signed-rank test statistic based on Tyler's angles
Q^2_{ICMS}	marginal sign test statistic in the symmetric IC model
Q^2_{ICMR}	marginal signed-rank test statistic in the symmetric IC model

1 Introduction

A location test is a statistical hypothesis test that compares the location of one sample to a given constant, or that compares the locations of several samples to each other. In this thesis, we are concerned with one-sample testing. In a multivariate context, we are dealing simultaneously with more than one variable. Multivariate data analysis is applied when several measurements or observations are made on several individuals or experimental units. Thus, we are naturally led to consider vector-valued observations. Multivariate analysis takes into account the statistical dependence between the variables of a data set. If each variable is studied separately, then some useful relationships among the variables may not be detected. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of i.i.d. d-variate observations and write $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for the corresponding $d \times n$ data matrix. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be the location parameter of interest. In the one-sample location problem, we wish to test whether $\boldsymbol{\theta}$ is equal to some fixed value $\boldsymbol{\theta}_0$:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

The literature proposes a vast list of multivariate one-sample location tests. The classical parametric procedure for the one-sample location problem is Hotelling's T^2 test (Hotelling, 1931) which assumes that the data comes from a multivariate normal distribution. The inference is based on the use of the sample mean vector and the sample covariance matrix. Hotelling's T^2 test is optimal under the assumption of normality. See, for example, the book by Anderson (2003). The test is, however, sensitive to outlying observations and poor in its efficiency for heavy-tailed distributions (Oja, 2010). The test may lead to unreliable results when the underlying distribution strongly deviates from the assumed model. The assumption of normality is also hard to justify in practice. As a reaction to this, many nonparametric methods have been developed in the literature with the objective of extending to the multivariate context the classical univariate sign and rank techniques. Nonparametric methods can be applied with few assumptions about the underlying distribution of the data. We consider multivariate locations tests based on the following multivariate notions of sign and rank: (i) marginal signs and ranks (see, for example, Puri and Sen, 1971), (ii) spatial signs and ranks (Oja, 2010), and (iii) Oja signs and ranks (Oja, 1999). The parametric Hotelling's T^2 test serves as a reference test. The goal is to provide through a simulation study practical guidelines which test might be most useful in practice.

A central requirement for multivariate tests and estimates is that they are invariant and equivariant, respectively, under affine transformations of the data (Nordhausen, Oja, and Tyler, 2006). An affine transformation of the data vector $\mathbf{x}_i$ is a transformation of the form

$$\mathbf{x}_i \to \mathbf{A}\mathbf{x}_i + \mathbf{b}, \qquad i = 1, \ldots, n$$

– or, equivalently,

$$\mathbf{X} \to \mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_n',$$

where $\mathbf{A}$ is a full-rank $d \times d$ matrix and $\mathbf{b}$ is a d-vector. The vector $\mathbf{1}_n$ is an n-vector full of ones. Affine invariance feature ensures that the results of a multivariate data analysis do not depend on the chosen coordinate system. This means, for example, that a decision in favor of or against H_0 should be the same for the original data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and any affine transformation $\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_n$ of that data. Estimates are called affine equivariant if any affine transformation of the data is paralled by a similar transformation of the estimate. All multivariate tests are not affine invariant automatically. The procedures have to be sometimes modified to obtain affine invariant test versions. Affine invariance can be achieved with appropriate data transformation techniques. Some of these techniques are discussed in Section 5.3.

The inference in multivariate analysis is usually based on location and scatter statistics. The following definitions summarize some important properties for multivariate estimators.

Definition 1.1. (Nordhausen et al., 2006) (i) A d-vector valued statistic $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is called a location statistic if it is affine equivariant, that is,

$$\mathbf{T}(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_n') = \mathbf{A}\mathbf{T}(\mathbf{X}) + \mathbf{b}$$

for all full-rank $d \times d$ matrices $\mathbf{A}$, and all d-vectors $\mathbf{b}$.

(ii) A $d \times d$ matrix $\mathbf{\Sigma} = \mathbf{\Sigma}(\mathbf{X}) \geq 0$ is a scatter statistic if it is affine equivariant in the sense that

$$\mathbf{\Sigma}(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_n') = \mathbf{A}\mathbf{\Sigma}(\mathbf{X})\mathbf{A}'$$

for all full-rank $d \times d$ matrices $\mathbf{A}$, and all d-vectors $\mathbf{b}$.

(iii) A scatter statistic with respect to the origin is affine equivariant in the sense that

$$\mathbf{\Sigma}(\mathbf{A}\mathbf{X}\mathbf{J}) = \mathbf{A}\mathbf{\Sigma}(\mathbf{X})\mathbf{A}'$$

for all full-rank $d \times d$ matrices $\mathbf{A}$, and for all $n \times n$ sign-change matrices $\mathbf{J}$ (a diagonal matrix with entries ± 1).

If $\mathbf{X}$ is a random sample, then it is natural to require that the statistics are invariant under permutations of the observations, that is,

$$\mathbf{T}(\mathbf{X}\mathbf{P}) = \mathbf{T}(\mathbf{X}) \quad \text{and} \quad \mathbf{\Sigma}(\mathbf{X}\mathbf{P}) = \mathbf{\Sigma}(\mathbf{X})$$

for all $n \times n$ permutation matrices $\mathbf{P}$ (obtained by permuting the rows or columns of $\mathbf{I}_d$).

The most common location and scatter statistics are the vector of means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. An important aim in multivariate statistics is to develop affine equivariant robust alternatives to the sample mean vector and the sample covariance matrix. Robustness in this context means that the methods should not be sensitive to the violations of the assumptions. Examples of robust nonparametric location estimates include the affine equivariant spatial median (Hettmansperger and Randles, 2002) and the Oja median (Oja, 1983). Examples of robust estimates of scatter matrix are the sign and rank covariance matrices (Visuri, Koivunen, and Oja, 2000). If affine equivariant signs and ranks are used, then the corresponding covariance matrices are affine equivariant. Another important affine equivariant estimator of scatter is the one proposed by Tyler (1987). Tyler's scatter matrix (with fixed location $\mathbf{T}(\mathbf{X})$) is defined as

$$\boldsymbol{\Sigma}(\mathbf{X}) = d \cdot \mathrm{E}\left[\frac{(\mathbf{X}-\mathbf{T}(\mathbf{X}))(\mathbf{X}-\mathbf{T}(\mathbf{X}))'}{\|\mathbf{X}-\mathbf{T}(\mathbf{X})\|^2_{\boldsymbol{\Sigma}(\mathbf{X})}}\right],$$

where $\|\mathbf{X}-\mathbf{T}(\mathbf{X})\|_{\boldsymbol{\Sigma}(\mathbf{X})} = [(\mathbf{X}-\mathbf{T}(\mathbf{X}))'\boldsymbol{\Sigma}(\mathbf{X})^{-1}(\mathbf{X}-\mathbf{T}(\mathbf{X}))]^{1/2}$ denotes the Mahalanobis distance between $\mathbf{X}$ and $\mathbf{T}(\mathbf{X})$ with respect to $\boldsymbol{\Sigma}(\mathbf{X})$. Tyler's scatter matrix is often used in inner standardizations when creating affine invariant spatial sign and signed-ranks tests (Section 5.4) and it plays an important role in the location testing procedures proposed by Hallin and Paindaveine (2002b) (Section 5.6). Location and scatter statistics are finite-sample versions of location and scatter functionals.

Symmetry arises as a major assumption on distributions in multivariate nonparametric location inference. The notion of location is well-defined only under symmetry. Also, symmetry is needed for the differents location tests to be comparable. Therefore, addressing multivariate symmetry in this context is necessary. Symmetry concepts relevant to this thesis are discussed in Chapter 2.

As the assumption of normality may not always hold in practice, it is important to find richer families of models which may include the normal distribution that are more flexible, but still analytically and computationally tractable. L_p-norm distributions offer an alternative to normally distributed data (Song and Gupta, 1997; Gupta and Song, 1997). These distributions are used, for example, in robustness studies. As extensions to normally distributed data, L_p-norm distributions are used in this thesis as simulation data to see how well the different tests perform when the underlying distribution deviates from normality.

In our simulation study, the powers of the proposed tests are investigated under various underlying distributions, sample sizes, dimensions, and deviations from the null value to determine under which settings different tests perform best. We compare in turn the powers of the nonparametric tests and the powers

of Hotelling's T^2 test relative to the powers of the nonparametric tests. Based on these studies, practical guidelines are given as to which test(s) one should use.

The outline of the thesis is as follows. In the next chapter, different multivariate symmetry concepts are discussed. In Chapter 3, L_p-norm distributions, their properties, and methods to generate random samples from them are presented. Chapter 4 introduces the different multivariate sign and rank concepts which form the basis of the nonparametric multivariate location tests. Chapter 5 presents the different tests and methods to attain affine invariant/ equivariant tests/estimates. The simulation study is conducted in Chapter 6. Finally, Chapter 7 concludes the thesis by summarizing the main findings and suggestions.

2 Multivariate symmetry concepts

Symmetry of multivariate distributions plays an important role in multivariate statistical inference. In the multivariate location problem, the notion of location is well-defined only under symmetry: when a distribution is symmetric about a given center in some sense, a natural location would be the center of symmetry. In addition, when some form of symmetry is required, the different tests and estimates then refer to the same population quantity. Serfling (2006) investigated various notions of symmetry and asymmetry. Here we introduce a few of the most common multivariate symmetry concepts. Existing multivariate symmetry concepts can be presented in various ways; for example, in terms of invariance of the distribution of a centered random vector $\mathbf{X} - \boldsymbol{\theta}$ in $\mathbb{R}^d$ under a suitable family of transformations or in terms of properties of the probability density function. The symmetry concepts are presented in an increasing order of generality: spherical symmetry implies elliptical symmetry which in turn implies exchange-sign symmetry, and further, exchange-sign symmetry implies central symmetry. All the symmetry concepts discussed reduce to the same notion of univariate symmetry. The notion of symmetry in the univariate case is straightforward: a random variable X is said to be symmetric about a given center $\theta \in \mathbb{R}$ if $X - \theta \sim -(X - \theta)$, where $\sim$ stands for equality in distribution.

2.1 Spherical symmetry

Spherical symmetry is the strongest assumption among traditional notions of multivariate symmetry. A random d-vector $\mathbf{X}$ has a distribution spherically symmetric about a point $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathbf{X} - \boldsymbol{\theta} \sim \mathbf{O}(\mathbf{X} - \boldsymbol{\theta})$ for all orthogonal $d \times d$ matrices $\mathbf{O}$. In other words, rotations and reflections have no impact on the distribution of $\mathbf{X}$ about $\boldsymbol{\theta}$. The density of a spherically distributed vector $\mathbf{x} \in \mathbb{R}^d$ is a function of $(\mathbf{x} - \boldsymbol{\theta})'(\mathbf{x} - \boldsymbol{\theta})$. All random vectors $\mathbf{X}$ spherically symmetric about $\boldsymbol{\theta}$ can be decomposed into $R\mathbf{U}$, where the random variable $R = \|\mathbf{X} - \boldsymbol{\theta}\|$ is the length of the random vector $\mathbf{X} - \boldsymbol{\theta}$ and $\mathbf{U} = \frac{\mathbf{X} - \boldsymbol{\theta}}{\|\mathbf{X} - \boldsymbol{\theta}\|}$ is its direction vector. $\mathbf{U}$ is uniformly distributed on the unit sphere $\mathcal{S}^d = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| = 1\}$ in $\mathbb{R}^d$. Moreover, R and $\mathbf{U}$ are independent. This way of presenting spherically symmetric distributions becomes relevant in Chapter 3 when we generalize spherical distributions to L_p-norm spherical distributions. However, L_p-norm spherical distributions are no longer spherically symmetric, but rather they are exchange-sign symmetric (Section 2.3).

2.2 Elliptical symmetry

A generalization of the concept of spherical symmetry is that of elliptical symmetry. A random d-vector $\mathbf{X}$ has an elliptically symmetric distribution with parameters $\boldsymbol{\theta}$ (center) and $\boldsymbol{\Sigma}$ (scatter matrix) if it is obtained as follows:

$$\mathbf{X} \sim \mathbf{A}\mathbf{Y} + \boldsymbol{\theta},$$

where the full-rank $d \times d$ matrix $\mathbf{A}$ satisfies $\mathbf{A}'\mathbf{A} = \boldsymbol{\Sigma}$, and $\mathbf{Y}$ is a d-variate random vector spherically symmetric about $\mathbf{0}$. In other words, an affine transformation of a spherically symmetric random vector generates an elliptically symmetric random vector. The family of elliptically symmetric distributions is thus closed under affine transformations. All elliptically symmetric distributions can also be transformed into spherically symmetric distributions. If $\mathbf{x} \in \mathbb{R}^d$ is an elliptically distributed vector, then its density is a function of $(\mathbf{x}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\theta})$. Optimal signed-rank score tests by Hallin and Paindaveine (2002a, 2002b) discussed in Section 5.6 assume strict ellipticity.

2.3 Exchange-sign symmetry

Exchange-sign symmetry is similiar to elliptical symmetry, but offers a broader range of densities. A random d-vector $\mathbf{X}$ is said to be sign-symmetric about a center $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathbf{X} - \boldsymbol{\theta} \sim \mathbf{J}(\mathbf{X} - \boldsymbol{\theta})$ for all sign-change matrices $\mathbf{J}$. This implies the symmetric independent component (IC) model (see Section 5.7). The distribution of $\mathbf{X}$ is exchange-sign symmetric about a point $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathbf{X} - \boldsymbol{\theta} \sim \mathbf{P}\mathbf{J}(\mathbf{X} - \boldsymbol{\theta})$ for all permutation matrices $\mathbf{P}$ and all sign-change matrices $\mathbf{J}$. The densities f of exchange-sign symmetric vectors $\mathbf{x} \in \mathbb{R}^d$ satisfy $f(\mathbf{x}) = f(\mathbf{P}\mathbf{J}\mathbf{x})$. Thus, the components of $\mathbf{x}$ are exchangeable and marginally symmetric around $\mathbf{0}$.

Generalizations of spherical distributions, L_p-norm spherical distributions $S(d, p)$, are exchange-sign symmetric. The L_p-norm uniform distribution $U(d, p)$ has an exchange-sign symmetric distribution on the L_p-norm unit sphere $\mathcal{S}_p^d = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_p = 1\}$ in $\mathbb{R}^d$, where $\|\mathbf{x}\|_p$ is an L_p-norm of $\mathbf{x}$. L_p-norm distributions are discussed in Chapter 3.

2.4 Central symmetry

Perhaps the most direct extension of the univariate concept of symmetry to the multivariate setting is central symmetry. A random d-vector $\mathbf{X}$ has a distribution centrally symmetric (or simply symmetric) about a point $\boldsymbol{\theta} \in \mathbb{R}^d$ if $\mathbf{X}-\boldsymbol{\theta} \sim -(\mathbf{X}-\boldsymbol{\theta})$. The density f of a centrally distributed vector $\mathbf{x} \in \mathbb{R}^d$ satisfies $f(\mathbf{x} - \boldsymbol{\theta}) = f(\boldsymbol{\theta} - \mathbf{x})$. Central symmetry is a broader concept of symmetry than the previous ones: spherically and elliptically symmetric distributions are also centrally symmetric. However, not all centrally symmetric random vectors are spherically or elliptically symmetric. For example, the uniform distribution

on the d-dimensional hypercube $[-1,1]^d = \{(z_1,\ldots,z_d)' \in \mathbb{R}^d \mid -1 \leq z_i \leq 1, i = 1,\ldots,d\}$ is centrally symmetric but not spherically or elliptically symmetric. Central symmetry is also a much weaker assumption than spherical or elliptical symmetry. Therefore, the corresponding location tests may be more robust than their spherical or elliptical competitors. Multivariate location testing procedures based on spatial signs and ranks (Section 5.3) and Oja signs and ranks (Section 5.4) only require the distribution of $\mathbf{X}$ to be centrally symmetric about $\boldsymbol{\theta}$.

The standard multivariate normal distribution and the multivariate t-distribution are symmetric in every sense described. The general multivariate normal distribution is not necessarily spherically symmetric but is elliptically symmetric. In this thesis, the simulation data come from distributions known to be at least exchange-sign symmetric about $\boldsymbol{\theta}$. In general, actual distributions are not typically symmetric in any strict sense. Rather, symmetric distributions are used only as approximations to actual distributions in modeling (Zuo, 1998).

3 L_p-norm distributions

3.1 General definition

L_p-norm (or p-norm) distributions are a class of multivariate probability distributions. These distributions can be used to generalize some familiar multivariate distributions like the standard multivariate normal distribution and the multivariate t-distribution. As the name suggests, L_p-norm distributions are based on L_p-norms. The L_p-norm function generalizes the regular Euclidean norm. Let $\mathbf{x} = (x_1, \ldots, x_d)' \in \mathbb{R}^d$ and $p > 0$. The L_p-norm of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}.$$

For $p = 2$, the regular Euclidean norm is obtained. We will hold to the naming convention of Gupta and Song (1997) and call $\|\mathbf{x}\|_p$ an L_p-norm even though the triangle inequality only holds for $p \geq 1$.

Most multivariate analysis techniques assume multivariate normal models. Natural data, however, deviate often significantly from normality. Therefore, it is important to find richer families of models that may include the normal distribution that are more flexible but still computationally and analytically tractable. L_p-norm distributions offer an alternative to normally distributed data. L_p-norm distributions include the L_p-norm uniform distribution and the L_p-norm spherical distribution. An overview of the L_p-norm uniform distribution can be found in Song and Gupta (1997) and one of L_p-norm spherical distributions in Gupta and Song (1997). The L_p-norm uniform distribution is a generalization of the (L_2-norm) uniform distribution and is used in constructing L_p-norm spherical distributions. Examples of L_p-norm spherical distributions include the p-generalized normal distribution and the L_p-norm multivariate t-distribution. They are generalizations of the multivariate normal distribution and the multivariate t-distribution, respectively. Definitions of L_p-norm uniform and L_p-norm spherical distributions are given next.

Definition 3.1. (Gupta and Song, 1997) The random vector $\mathbf{U}_d = (U_1, \ldots, U_d)'$ is said to have an L_p-norm uniform distribution, denoted by $U(d, p)$, if

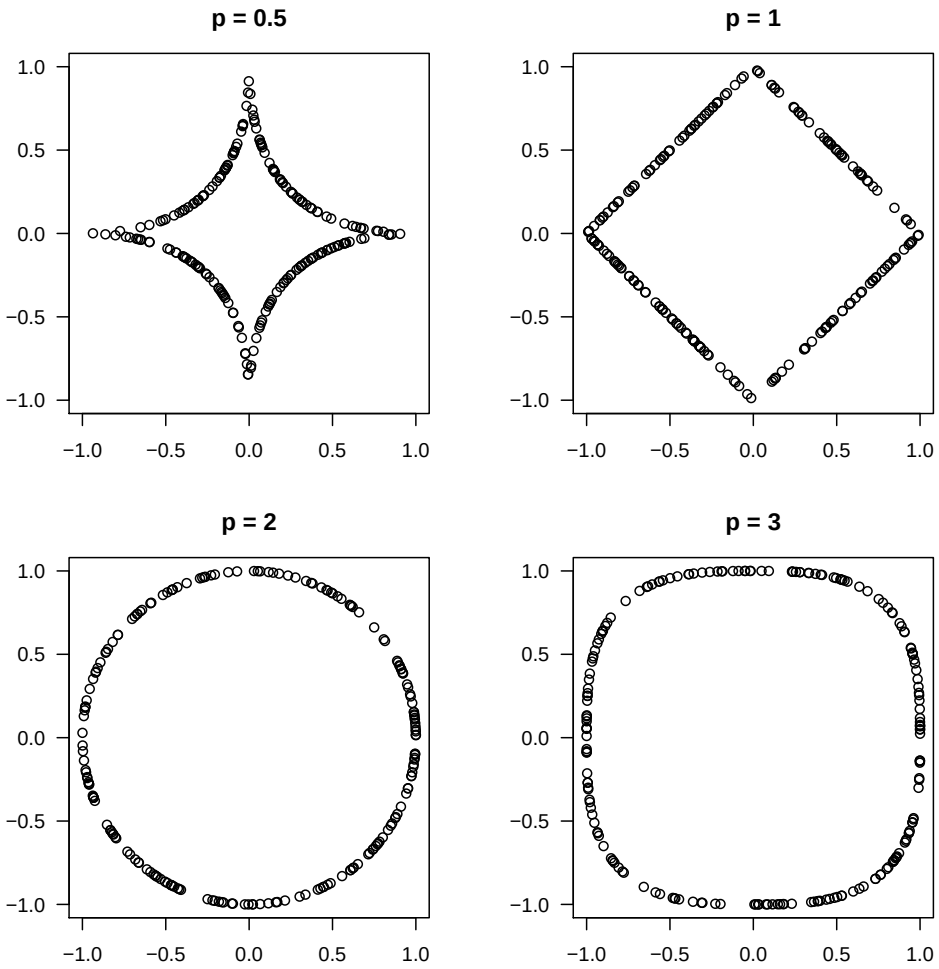

Figure 3.1. Random samples of size $n = 200$ from the L_p-norm uniform distribution $U(2, p)$ (distributed on the L_p-norm unit spheres $\mathcal{S}_p^2$) for some selected values of p.

$\sum_{i=1}^{d} |U_i|^p = 1$ and the joint p.d.f. of $U_1, \ldots, U_{d-1}$ is given by

$$f(u_1, \ldots, u_{d-1}) = \frac{p^{d-1}\Gamma(d/p)}{2^{d-1}\Gamma^d(1/p)} \left(1 - \sum_{i=1}^{d-1} |u_i|^p\right)^{(1-p)/p},$$
$$-1 < u_i < 1,\ i = 1, \ldots, d-1,\ \sum_{i=1}^{d-1} |u_i|^p < 1,$$

where $\Gamma(\cdot)$ denotes the gamma function. For $p = 2$, $U(d, p)$ becomes the regular uniform distribution. L_p-norm uniform distributions are exchange-sign symmetric and said to be uniformly distributed on the surface of the L_p-norm unit sphere $\mathcal{S}_p^d = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_p = 1\}$ in $\mathbb{R}^d$. The L_p-norm unit sphere is a generalization of the regular (L_2-norm) unit sphere. Figure 3.1 illustrates some randoms samples of size $n = 200$ from the L_p-norm uniform distribution $U(2, p)$ (distributed on the L_p-norm unit spheres $\mathcal{S}_p^2$) for some selected values of p. As

$p \to 0$, the shape of the distribution becomes more and more "cross-like", and as $p \to \infty$, the distribution begins to form the shape of a (unit) square. An algorithm for generating random samples from the L_p-norm uniform distribution (according to Liang and Ng, 2008) is presented in Section 3.3.1.

Definition 3.2. (Gupta and Song, 1997) A d-variate random vector $\mathbf{X}$ is said to have an L_p-norm spherical distribution (denoted by $\mathbf{X} \sim S(d,p)$) if $\mathbf{X} = R\mathbf{U}_d$, where $\mathbf{U}_d \sim U(d,p)$ and R, which is independent of $\mathbf{U}_d$, is a univariate nonnegative random variable with c.d.f. $F(\cdot)$.

For $p = 2$, $S(d,p)$ becomes the regular spherical distribution S_d. L_p-norm spherical distributions are of great practical interest since they offer more flexibility than the spherically symmetric model.

3.2 Some properties

We will now take a closer look at the properties of these distributions. The contents of this section is derived from Gupta and Song (1997).

The L_p-norm uniform distribution is a special case of the L_p-norm spherical distribution: if $P(R = 1) = 1$, then $\mathbf{X} \sim \mathbf{U}_d$ and therefore $\mathbf{U}_d \sim S(d,p)$. If $\mathbf{X} \sim S(d,p)$ with $P(\mathbf{X} = \mathbf{0}) = 0$, then the distribution of $\mathbf{X}$ will be denoted by $S^+(d,p)$. If $\mathbf{X} = R\mathbf{U}_d \sim S^+(d,p)$, then $\|\mathbf{X}\|_p \sim R$ and $\mathbf{X}/\|\mathbf{X}\|_p \sim \mathbf{U}_d$. Thus, R ("radius") is the length of $\mathbf{X}$ (in the L_p-norm sense) and $\mathbf{U}_d$ is its direction vector. In addition, $\|\mathbf{X}\|_p$ and $\mathbf{X}/\|\mathbf{X}\|_p$ are independent. The densities of L_p-norm spherically distributed vectors $\mathbf{x} \in \mathbb{R}^d$ are a function of $\|\mathbf{x}\|_p$.

Different L_p-norm spherical distributions can be generated from L_p-norm uniform distributions by multiplying them with different positive random variables R. The distribution of R is uniquely defined by the distribution of $\mathbf{X} \sim S(d,p)$. Random variables R have a general probability density form

$$f(r) = \frac{2^d \Gamma^d(1/p)}{p^{d-1}\Gamma(d/p)} r^{d-1} g(r^p), \qquad r > 0,$$

where the function $g(\cdot)$ is called the density generator of $S(d,p)$ and $f(\cdot)$ is the generating density of the distribution. Different $S(d,p)$ distributions have different density generator functions for R. In regard to this thesis, some important subclasses of L_p-norm spherical distributions are the aforementioned p-generalized normal distribution and the L_p-norm multivariate t-distribution.

Following the notation of Gupta and Song (1997), we let $N_d(\mathbf{0}, \mathbf{I}_d, p)$ denote the d-variate p-generalized normal distribution. The random vector $\mathbf{X} \sim N_d(\mathbf{0}, \mathbf{I}_d, p)$ has the probability density function

$$f(x_1, \ldots, x_d) = \frac{p^d r_0^{d/p}}{2^d \Gamma^d(1/p)} e^{-r_0 \sum_{i=1}^d |x_i|^p}, \quad -\infty < x_i < \infty,\ i = 1, \ldots, d,$$

where r_0 is a parameter. We know that $\mathbf{X}$ can be decomposed into $R\mathbf{U}_d$, where

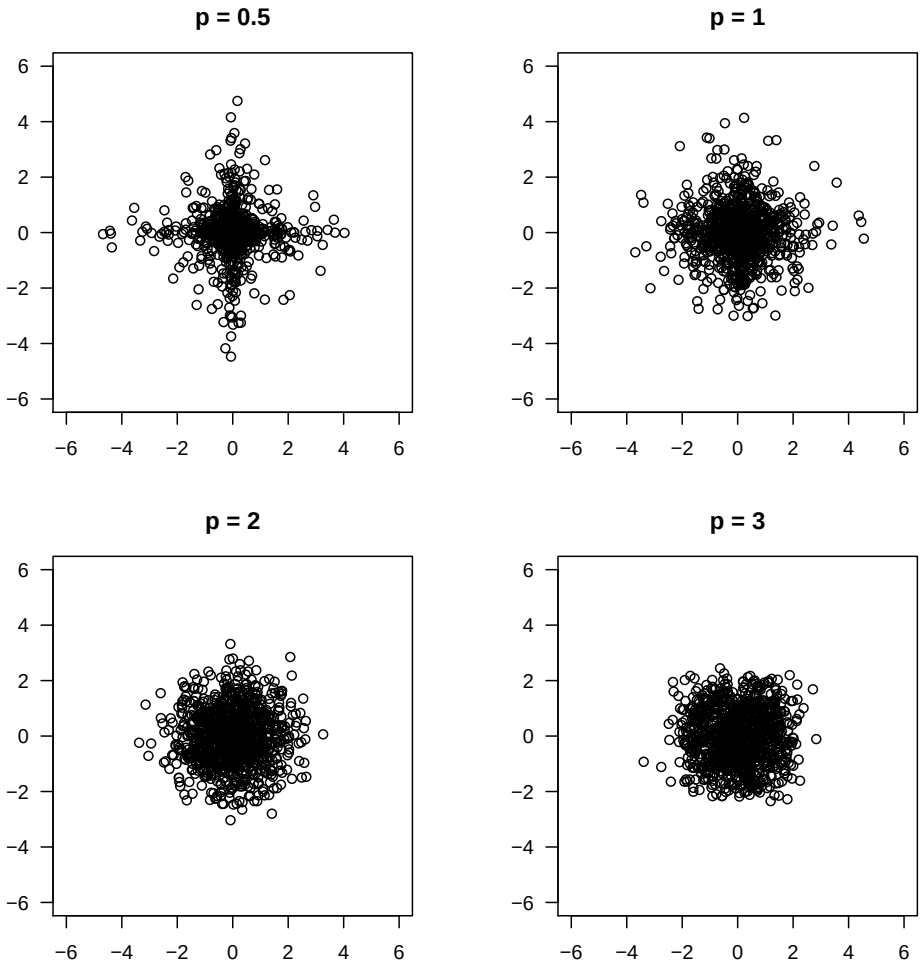

Figure 3.2. Random samples of size $n = 1000$ from the p-generalized normal distribution $N_2(\mathbf{0}, \mathbf{I}_2, p)$ for some selected values of p.

$\mathbf{U}_d \sim U(d, p)$, and R has the probability density function

$$f(r) = \frac{p r_0^{d/p}}{\Gamma(d/p)} r^{d-1} e^{-r_0 r^p}, \qquad r > 0.$$

For $p = 2$, the p-generalized normal distribution reduces to the multivariate normal distribution. L_p-norm spherical distributions with $p \neq 2$ are no longer invariant with respect to rotations, that is, they are no longer spherically symmetric. Instead, they are only exchange-sign symmetric. Figure 3.2 illustrates some random samples of size $n = 1000$ from the p-generalized normal distribution $N_2(\mathbf{0}, \mathbf{I}_2, p)$ for some selected values of p. The shapes of the L_p-norm uniform distributions are visible. The values of R spread the data points. Figure 3.3 illustrates a kernel density estimation of the distribution of a p-generalized normal distribution $N_2(\mathbf{0}, \mathbf{I}_2, p)$ based on a random sample of size $n = 1000$ for some selected values of p. Kernel density estimation is a nonparametric met-

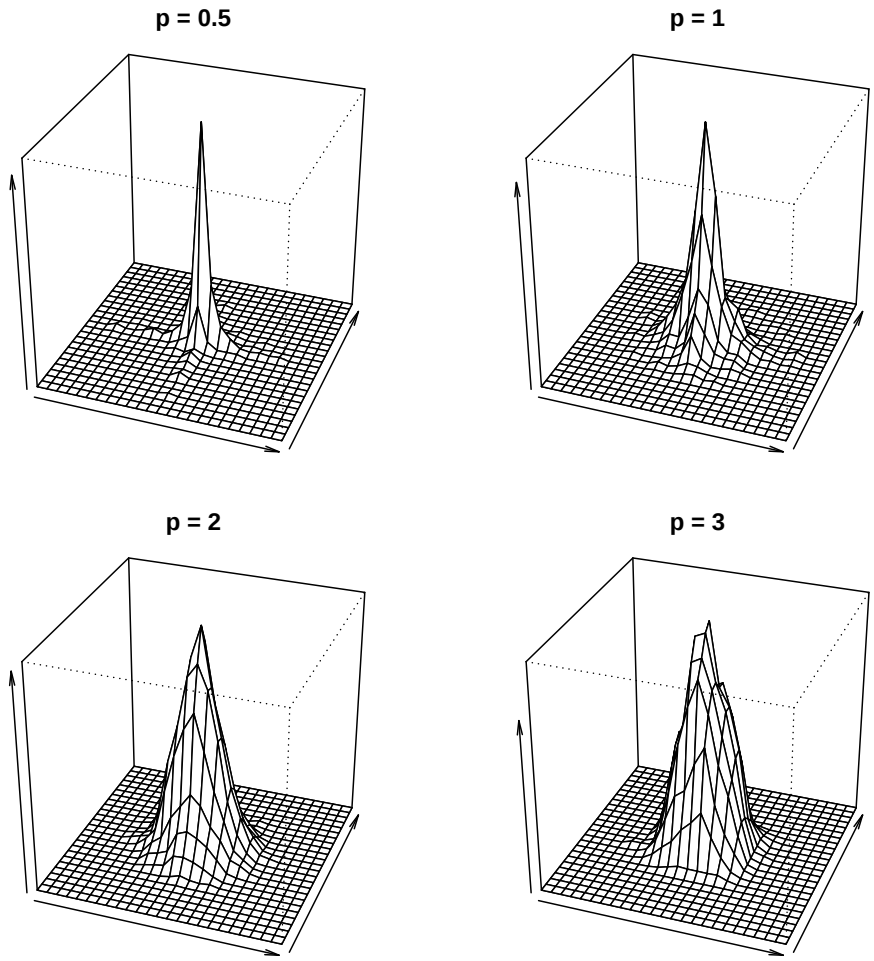

Figure 3.3. A kernel density estimation of the distribution of a p-generalized normal distribution $N_2(\mathbf{0}, \mathbf{I}_2, p)$ based on a random sample of size $n = 1000$ for some selected values of p.

hod to estimate the probability density functions of random variables. The method was independently developed by Rosenblatt (1956) and Parzen (1962). An algorithm for generating random samples from the p-generalized normal distribution (according to Liang and Ng, 2008) is presented in Section 3.3.2.

If a d-variate random vector $\mathbf{X}$ follows the L_p-norm multivariate t-distribution, then its probability density function is of the form

$$f(x_1, \ldots, x_d) = \frac{p^d \Gamma((d+df)/2) s^{-d/p}}{2^d \Gamma^d(1/p) \Gamma((d+df)/2 - d/p)} \left(1 + \left(\sum_{i=1}^{d} |x_i|^p\right) \Big/ s\right)^{-(d+df)/2},$$

$$-\infty < x_i < \infty,\ df > 0 \text{ an integer},\ (d+df)/2 > d/p,\ s > 0.$$

For $p = 2$, the multivariate t-distribution is obtained. Again, $\mathbf{X}$ can be decom-

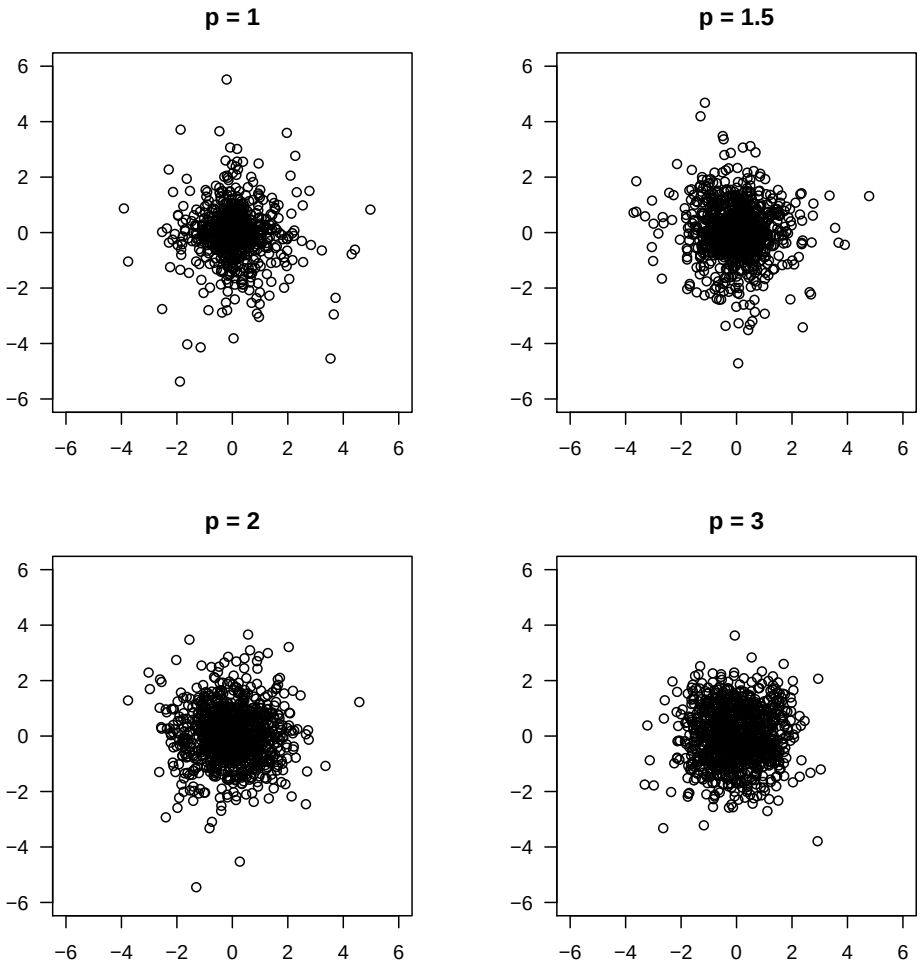

Figure 3.4. Random samples of size $n = 1000$ from a 2-variate L_p-norm multivariate t-distribution with mean $\mathbf{0}$, covariance $\mathbf{I}_2$, and 9 degrees of freedom for some selected values of p.

posed into $R\mathbf{U}_d$, where $\mathbf{U}_d \sim U(d, p)$, and R has the probability density function

$$f(r) = \frac{p\Gamma((d+df)/2)}{\Gamma(d/p)\Gamma((d+df)/2 - d/p)} r^{d-1} s^{-d/p} (1 + r^p/s)^{-(d+df)/2}, \quad r > 0.$$

Figure 3.4 illustrates some random samples of size $n = 1000$ from a 2-variate L_p-norm multivariate t-distribution with mean $\mathbf{0}$, covariance $\mathbf{I}_2$, and 9 degrees of freedom for some selected values of p. Relative to p-generalized normal distributions, the data points are more spread out. (A different set of p-values is used in the figure due to the properties of the probability density functions of L_p-norm multivariate t-distributions which require that $(d + df)/2 > d/p$. Also, the integral function needed to compute the covariance matrix of L_p-norm multivariate t-distributions is convergent only for these value combinations of p and df. The use of 9 degrees of freedom allows 4 different values of p to be

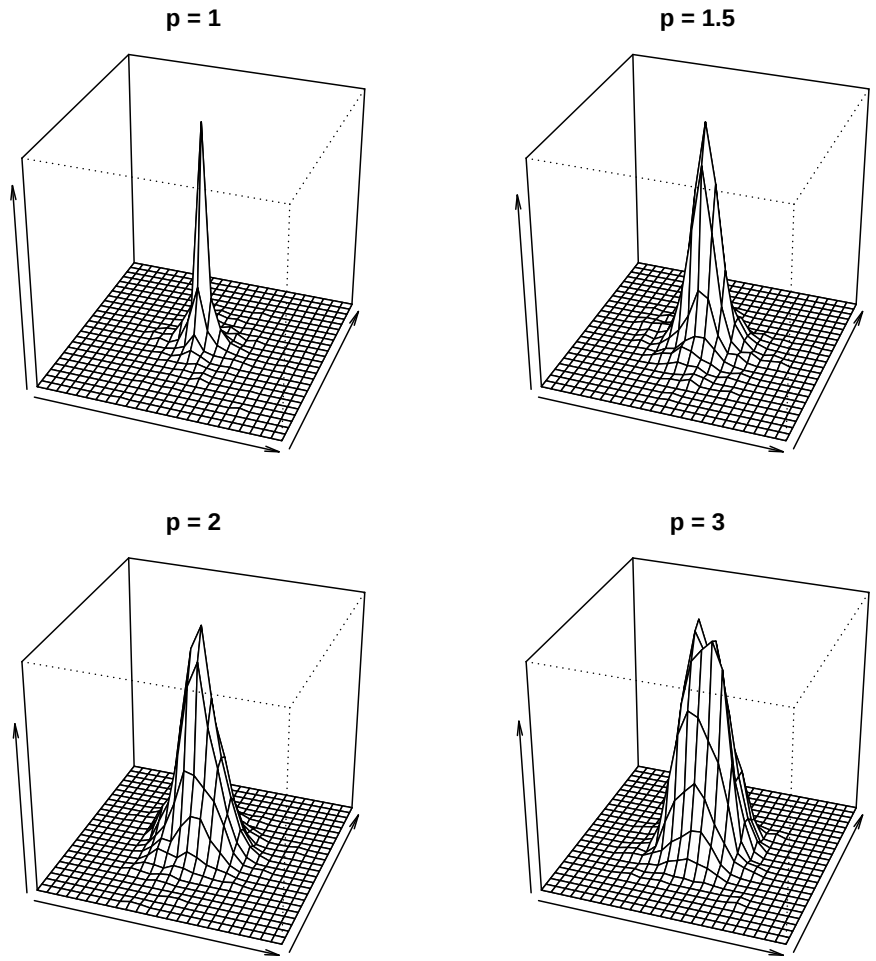

Figure 3.5. A kernel density estimation of the distribution of a 2-variate L_p-norm multivariate t-distribution with mean $\mathbf{0}$, covariance $\mathbf{I}_2$, and 9 degrees of freedom based on a random sample of size $n = 1000$ for some selected values of p.

used.) Figure 3.5 illustrates a kernel density estimation of the distribution of a 2-variate L_p-norm multivariate t-distribution with mean $\mathbf{0}$, covariance $\mathbf{I}_2$, and 9 degrees of freedom based on a random sample of size $n = 1000$ for some selected values of p. An algorithm for generating random samples from the L_p-norm multivariate t-distribution using the acceptance-rejection method (von Neumann, 1951) is presented in Section 3.3.3.

Next, we consider the expected value and covariance properties of these distributions. Suppose $\mathbf{X} = R\mathbf{U}_d \sim S(d, p)$. Then (Gupta and Song, 1997)

$$\mathrm{E}(\mathbf{X}) = \mathbf{0}$$

and the covariance matrix, provided it exists, has a general form

$$\mathrm{cov}(\mathbf{X}) = \frac{\Gamma(d/p)\Gamma(3/p)}{\Gamma(1/p)\Gamma((d+2)/p)}\mathrm{E}(R^2)\mathbf{I}_d.$$

Hence, the L_p-norm uniform distribution, the p-generalized normal distribution, and the L_p-norm multivariate t-distribution all have expected values of $\mathbf{0}$. To obtain the covariance matrices, one has to evaluate $\mathrm{E}(R^2)$ for each distribution. Since $R = 1$ for $\mathbf{U}_d \sim U(d, p)$,

$$\mathrm{cov}(\mathbf{U}_d) = \frac{\Gamma(d/p)\Gamma(3/p)}{\Gamma(1/p)\Gamma((d+2)/p)}\mathbf{I}_d.$$

Let $\mathbf{X} \sim N_d(\mathbf{0}, \mathbf{I}_d, p)$. Now

$$\mathrm{E}(R^2) = \frac{p^{2/p}\Gamma((d+2)/p)}{\Gamma(d/p)}$$

and therefore

$$\mathrm{cov}(\mathbf{X}) = \frac{p^{2/p}\Gamma(3/p)}{\Gamma(1/p)}\mathbf{I}_d.$$

Finally, let $\mathbf{X}$ follow the L_p-norm multivariate t-distribution. Now

$$\mathrm{E}(R^2) = \int_0^\infty r^2 f(r)dr = \int_0^\infty \frac{p\Gamma((d+df)/2)r^{d+1}}{\Gamma(d/p)\Gamma((d+df)/2 - d/p)} s^{-d/p}(1 + r^p/s)^{-(d+m)/2}dr$$

has no simple closed form. The value of $\mathrm{E}(R^2)$ and the resulting value of $\mathrm{cov}(\mathbf{X})$ will be calculated numerically for the simulation study.

3.3 Generating random samples

We want to generate random samples from (i) the L_p-norm uniform distribution and from L_p-norm spherical distributions (ii) p-generalized normal distribution and (iii) L_p-norm multivariate t-distribution. Since all L_p-norm spherical distributions are a product of a random vector following an L_p-norm uniform distribution $U(d, p)$ and a nonnegative random variable R, it is relatively easy to generate random samples from L_p-norm spherical distributions since it only requires generating samples from the univariate distribution of R. Algorithms for generating random samples from L_p-norm uniform distributions and p-generalized normal distributions were presented by Liang and Ng (2008). For generating random samples from the radius of an L_p-norm multivariate t-distribution, we use the acceptance-rejection method (von Neumann, 1951).

3.3.1 Uniform distribution on the L_p-norm unit sphere

Liang and Ng (2008) proposed a method based on inverse transformation for generating uniformly scattered points $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ on the L_p-norm unit sphere $\mathcal{S}_p^d$. The method can easily be implemented in practice: let $\mathbf{u}_i = (u_{i1}, \ldots, u_{id})' \sim U(d, p),\ i = 1, \ldots, n$. $\mathbf{u}_i$ has a stochastic representation

$$\mathbf{u}_i \sim \mathbf{x}_i = (x_{i1}, \ldots, x_{id})',$$

where the components $x_{i1}, \ldots, x_{id}$ are given by

$$\begin{aligned}
x_{i1} &= \pm b_1^{1/p}, \\
x_{i2} &= \pm[(1 - |x_1|^p) b_2]^{1/p}, \\
&\vdots \\
x_{i(d-1)} &= \pm \left[\left(1 - \sum_{1=1}^{d-2} |x_i|^p \right) b_{d-1} \right]^{1/p}, \\
x_{id} &= \pm \left(1 - \sum_{1=1}^{d-1} |x_i|^p \right)^{1/p},
\end{aligned}$$

where $b_1, \ldots, b_{d-1}$ are independent and $b_k \sim \text{Beta}[1/p, (d - k)/p]$, $k = 1, \ldots, d - 1$.

3.3.2 p-generalized normal distribution

A method to generate random samples from the p-generalized normal distribution was also proposed by Liang and Ng (2008). If the random variable R has the density of the radius of a p-generalized normal distribution, then the random variable $Y = r_0 R^p$ with $r_0 = 1/p$ has the density

$$f(y) = \frac{1}{\Gamma(d/p)} y^{d/p-1} e^{-y}, \qquad y > 0.$$

In other words, Y follows a gamma distribution with parameters $k = d/p$ and $\theta = 1$. An i.i.d. sample $\{y_1, \ldots, y_n\}$ can easily be generated from the gamma distribution. Then, an i.i.d. sample $\{r_1, \ldots, r_n\}$ from R can be obtained by

$$r_i = (y_i / r_0)^{1/p}, \qquad i = 1, \ldots, n.$$

Finally, a random sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from $N_d(\mathbf{0}, \mathbf{I}_d, p)$ is obtained by

$$\mathbf{x}_i = r_i \mathbf{u}_i, \qquad i = 1, \ldots, n,$$

where $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is a random sample from $U(d, p)$.

3.3.3 L_p-norm multivariate t-distribution

In absence of a more straightforward solution, random samples from the radius of an L_p-norm multivariate t-distribution are generated using the acceptance-rejection method (von Neumann, 1951). The method is used in situations where we want to generate random samples from a density f on some set X, but lack a direct method of doing so. Let g be a density on X from which we know how to generate samples with the property that

$$\forall x \in X\colon f(x) \leq cg(x)$$

for some constant c. The algorithm goes as follows:

1. Generate x from distribution g.
2. Generate u from Unif[0,1].
3. If $u \leq f(x)/(cg(x))$
   return x.
   Otherwise
   return to step 1.

There are an infinite number of densities g and constants c that can be used. The only difference between them is computation time. The greater the distance between $f(x)$ and $cg(x)$, the more candidate draws will be rejected. If $f(x) = cg(x)$, then all draws will be accepted. We use $c = 5$ and the F-distribution with parameters 4 and 2 as the distribution g. An exact proof for the validity of these choices for all $x > 0$ is missing. The choices are based on graphical and computational evaluations which showed that $f(x) \leq cg(x)$ is satisfied up to large values of x ($x = 1,000,000$). Thus, the probability of violating the assumption of the method is small. With these choices, the number of iterations needed to draw a random sample of size n is always roughly $c \cdot n$.

This way we obtain an i.i.d. sample $\{r_1, \ldots, r_n\}$ from the radius R of an L_p-norm multivariate t-distribution. Then, as before, a random sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from the L_p-norm multivariate t-distribution is obtained by

$$\mathbf{x}_i = r_i \mathbf{u}_i, \qquad i = 1, \ldots, n,$$

where $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is a random sample from $U(d, p)$.

4 Multivariate concepts of sign and rank

4.1 Introduction

Replacing the observations of a data set by their signs or ranks is a common procedure in nonparametric statistics. This means in general loosing efficiency under normality, but one can hope to get robust and distribution-free methods. Before extending to the multivariate case, we will review the univariate sign, rank, and signed-rank concepts. Let $x_1, \ldots, x_n$ be a univariate data set. The univariate sign function is

$$S(x) = \begin{cases} +1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0, \end{cases}$$

the (centered) rank function is

$$R(x) = \frac{1}{n} \sum_{j=1}^{n} S(x - x_j),$$

and the signed-rank function is

$$Q(x) = \frac{1}{2n} \sum_{j=1}^{n} [S(x - x_j) + S(x + x_j)] = \frac{1}{2}[R(x) - R(-x)].$$

The numbers $S_i = S(x_i)$, $R_i = R(x_i)$, and $Q_i = Q(x_i)$, $i = 1, \ldots, n$, are the observed signs, the observed (centered) ranks, and the observed signed-ranks. Univariate signs and ranks form the basis of univariate nonparametric tests such as the sign test and the Wilcoxon signed-rank test, among others. Few assumptions about the underlying distribution of the data are needed. Univariate signs and ranks are linked with the possibility to order the data. In the multivariate context, however, the concept of ordering becomes much more complicated since there is no natural way of ordering the data points in a d-dimensional space. Many alternative multivariate extensions of the univariate sign and rank concepts have been proposed in the literature. Different generalized notions of sign and rank allow for different extensions of the univariate sign and rank procedures to be made. In this thesis, we discuss and

apply location tests based on the following multivariate notions of sign and rank: (i) marginal signs and ranks (Puri and Sen, 1971), (ii) spatial signs and ranks (Oja, 2010), and (iii) Oja signs and ranks (Oja, 1999). We now turn to presenting each of these. The corresponding tests are presented in Chapter 5 and implemented in Chapter 6.

4.2 Marginal signs and ranks

The most obvious extension of univariate signs and ranks is their componentwise application to a multivariate data set, leading to componentwise or marginal signs and ranks. A comprehensive overview of marginal signs and ranks and related test procedures can be found in Puri and Sen (1971). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a d-variate data set. The vector of marginal signs of the ith observation is

$$\mathbf{S}_1(\mathbf{x}_i) = (S(x_{i1}), \ldots, S(x_{id}))',$$

where $S(\cdot)$ denotes the (univariate) sign function, the marginal rank vector of the ith observation is

$$\mathbf{R}_1(\mathbf{x}_i) = (R(x_{i1}), \ldots, R(x_{id}))',$$

where $R(x_{ij})$ denotes the marginal rank of $|x_{ij}|$ among all $|x_{1j}|, \ldots, |x_{nj}|$, and the marginal signed-rank vector of the ith observation is

$$\mathbf{Q}_1(\mathbf{x}_i) = (Q(x_{i1}), \ldots, Q(x_{id}))',$$

where $Q(x_{ij})$ denotes the marginal signed-rank of $|x_{ij}|$ among all $|x_{1j}|, \ldots, |x_{nj}|$. These vectors are used to construct multivariate analogues of the univariate sign and rank tests. Marginal sign- and rank-based multivariate location tests are presented in Sections 5.3 and 5.7.

4.3 Spatial signs and ranks

Another multivariate extension of the univariate sign and rank concepts is spatial signs and ranks. The book by Oja (2010) gives a thorough review of spatial signs and ranks and corresponding test procedures. Let $\mathbf{x} \in \mathbb{R}^d$. The spatial sign, spatial rank and spatial signed-rank functions $\mathbf{S}_2(\mathbf{x}), \mathbf{R}_2(\mathbf{x})$, and $\mathbf{Q}_2(\mathbf{x})$ are defined as

$$\begin{aligned}
\mathbf{S}_2(\mathbf{x}) &= \begin{cases} \|\mathbf{x}\|^{-1}\mathbf{x}, & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0}, & \text{if } \mathbf{x} = \mathbf{0}, \end{cases} \\
\mathbf{R}_2(\mathbf{x}) &= \mathbf{AVE}\{\mathbf{S}_2(\mathbf{x} - \mathbf{x}_j)\}, \text{ and} \\
\mathbf{Q}_2(\mathbf{x}) &= \frac{1}{2}[\mathbf{R}_2(\mathbf{x}) - \mathbf{R}_2(-\mathbf{x})].
\end{aligned}$$

In the univariate case, regular sign, rank, and signed-rank functions are obtained. The sample spatial sign, rank, and signed-rank vectors are $\mathbf{S}_2(\mathbf{x}_i)$, $\mathbf{R}_2(\mathbf{x}_i)$, and $\mathbf{Q}_2(\mathbf{x}_i)$, $i = 1, \ldots, n$, respectively. The spatial sign $\mathbf{S}_2(\mathbf{x}_i)$ is a unit vector pointing in the direction of $\mathbf{x}_i$ whenever $\mathbf{x}_i \neq \mathbf{0}$. The spatial rank $\mathbf{R}_2(\mathbf{x}_i)$ (spatial signed-rank $\mathbf{Q}_2(\mathbf{x}_i)$) is a vector inside a unit sphere in $\mathbb{R}^d$, the direction and length of which roughly reflect the direction and length of the observation $\mathbf{x}_i$. Spatial signs, ranks, and signed-ranks are not affine equivariant. They are only orthogonal equivariant and the centered ranks are invariant under location shifts.

Spatial signs, ranks, and signed-ranks are used for testing and estimation in multivariate location problems. The spatial sign and signed-rank tests presented in Section 5.3 are based on spatial signs and signed-ranks. The optimal signed-rank score tests by Hallin and Paindaveine (2002b) presented in Section 5.6 are based on standardized spatial signs.

4.4 Oja signs and ranks

A third generalization of the univariate sign and rank concepts is Oja signs and ranks. For an overview of Oja signs and ranks and related test procedures, see Oja (1999). Oja signs and ranks are based on the so-called Oja median (Oja, 1983). The Oja median is in turn based on distances measured via volumes of simplices. The volume of the d-variate simplex determined by the $d+1$ vertices $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d}, \mathbf{x}$ is

$$V(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d}, \mathbf{x}) = \frac{1}{d!} \text{abs} \left\{ \det \begin{pmatrix} 1 & \cdots & 1 & 1 \\ \mathbf{x}_{i_1} & \cdots & \mathbf{x}_{i_d} & \mathbf{x} \end{pmatrix} \right\},$$

where abs and det denote absolute value and determinant, respectively. This volume forms the basis of the Oja objective function, which in turn produces the affine equivariant Oja median and the affine invariant Oja sign and signed-rank tests. Consider the objective functions

$$D_1(\mathbf{x}) = \mathbf{AVE} \left\{ \text{abs} \left\{ \det \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ \mathbf{0} & \mathbf{x}_{i_1} & \cdots & \mathbf{x}_{i_{d-1}} & \mathbf{x} \end{pmatrix} \right\} \right\}$$

and

$$D_2(\mathbf{x}) = \mathbf{AVE} \left\{ \text{abs} \left\{ \det \begin{pmatrix} 1 & \cdots & 1 & 1 \\ \mathbf{x}_{i_1} & \cdots & \mathbf{x}_{i_d} & \mathbf{x} \end{pmatrix} \right\} \right\}.$$

The Oja sign and rank functions, $\mathbf{S}_3(\mathbf{x})$ and $\mathbf{R}_3(\mathbf{x})$, are defined as the gradient functions of $D_1(\mathbf{x})$ and $D_2(\mathbf{x})$, that is,

$$\mathbf{S}_3(\mathbf{x}) = \nabla D_1(\mathbf{x}) \quad \text{and} \quad \mathbf{R}_3(\mathbf{x}) = \nabla D_2(\mathbf{x}).$$

$\mathbf{Q}_3(\mathbf{x}) = \mathbf{R}_{3;2n}(\mathbf{x})$ is defined for the multivariate signed-rank function, where $\mathbf{R}_{3;2n}(\mathbf{x})$ is the rank function computed on the combined data

$\{\mathbf{x}_1, \ldots, \mathbf{x}_n, -\mathbf{x}_1, \ldots, -\mathbf{x}_n\}$ including the original observations and their reflections. In the univariate case, the usual univariate sign, rank, and signed-rank functions are obtained. The sample Oja sign, rank, and signed-rank vectors are $\mathbf{S}_3(\mathbf{x}_i)$, $\mathbf{R}_3(\mathbf{x}_i)$, and $\mathbf{Q}_3(\mathbf{x}_i)$, $i = 1, \ldots, n$, respectively. $\mathbf{S}_3(\mathbf{x}_i)$ depends on $\mathbf{x}_i$ only through its direction $\|\mathbf{x}_i\|^{-1}\mathbf{x}_i$ and it roughly points to the direction of $\mathbf{x}_i$. The vector $-\mathbf{R}_3(\mathbf{x}_i)$, if originated from $\mathbf{x}_i$, points in the direction of the mass of the data and the length of $\mathbf{R}_3(\mathbf{x}_i)$ increases as $\mathbf{x}_i$ moves out from the center. $\mathbf{Q}_3(\mathbf{x}_i)$, if located at $\mathbf{x}_i$, points in the direction of the origin, and its length increases with $\|\mathbf{x}_i\|$. $\mathbf{S}_3(\mathbf{x}_i)$ and $\mathbf{R}_3(\mathbf{x}_i)$ are affine equivariant in the sense that if the signs $\mathbf{S}_3^*(\mathbf{x}_i)$ and ranks $\mathbf{R}_3^*(\mathbf{x}_i)$ are calculated from the transformed observations $\mathbf{x}_i^* = \mathbf{A}\mathbf{x}_i + \mathbf{b}$, $i = 1, \ldots, n$, with a full-rank $d \times d$ matrix $\mathbf{A}$ and a d-vector $\mathbf{b}$, then

$$\mathbf{S}_3^*(\mathbf{x}_i^*) = \mathbf{A}^*\mathbf{S}_3(\mathbf{x}_i) \quad \text{and} \quad \mathbf{R}_3^*(\mathbf{x}_i^*) = \mathbf{A}^*\mathbf{R}_3(\mathbf{x}_i),$$

where $\mathbf{A}^* = |det(\mathbf{A})|(\mathbf{A}^{-1})'$.

Oja signs and ranks are used to construct multivariate analogues of the univariate sign and rank tests. Tests based on Oja signs and ranks are presented in Section 5.5.

Overall, multivariate signs and ranks are conceptually simple, natural generalizations of their univariate counterparts. For illustrative purposes, in Figure 4.1, one can see scatterplots for 50 bivariate observations from $N_2(\mathbf{0}, \mathbf{I}_2)$ with scatterplots for corresponding observed marginal, spatial, and Oja signs, ranks, and signed-ranks.

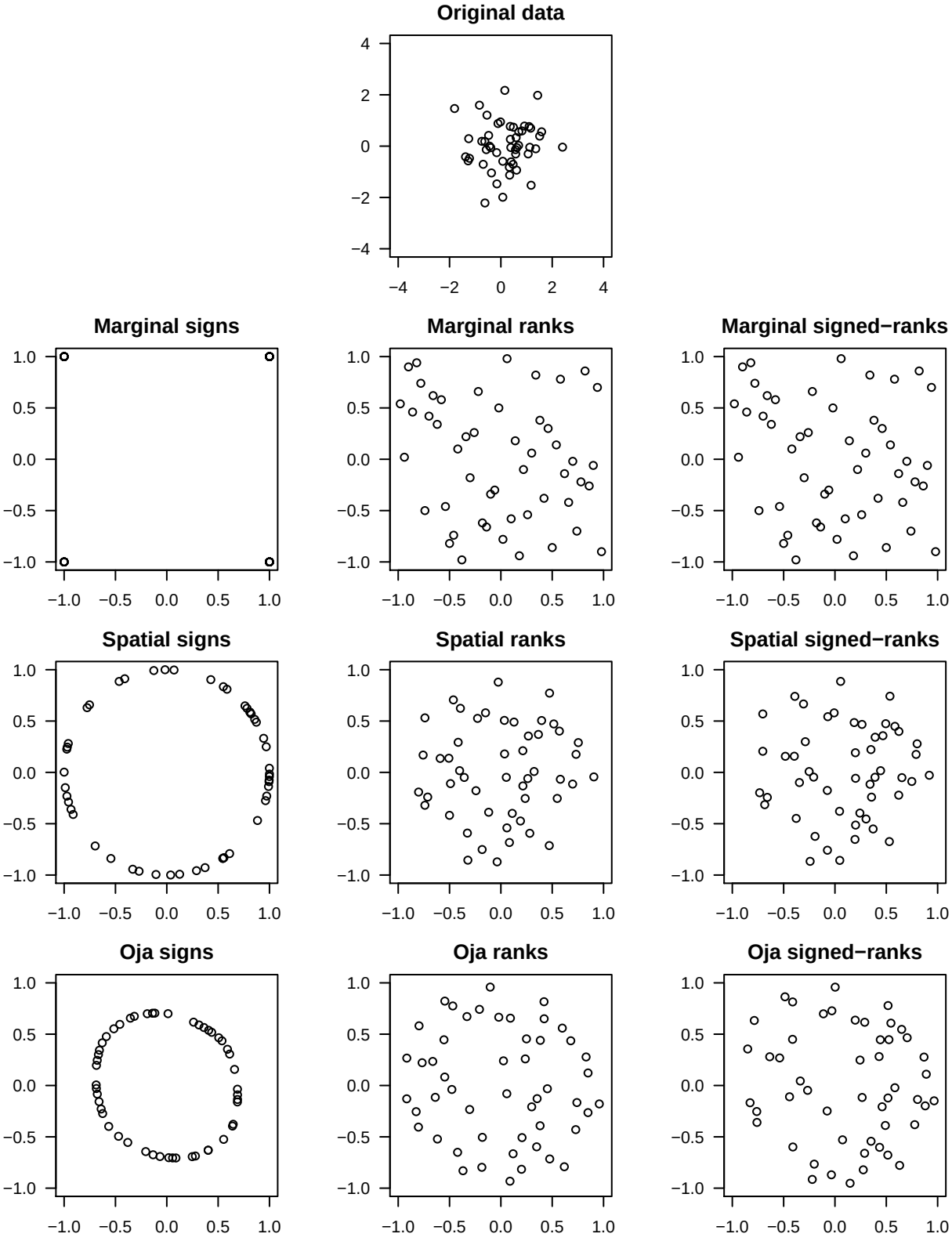

Figure 4.1. The scatterplots for a random sample of size $n = 50$ from $N_2(\mathbf{0}, \mathbf{I}_2)$ with scatterplots for corresponding observed marginal, spatial, and Oja signs, ranks, and signed-ranks.

5 Multivariate tests of location

5.1 Introduction

In this chapter, the different tests for the multivariate one-sample location problem are presented. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the $d \times n$ data matrix of n i.i.d. d-variate observations, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})'$, $i = 1, \ldots, n$. The test statistics are given for the hypotheses

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \mathbf{0}.$$

Any other null value $\boldsymbol{\theta}_0$ can be tested by replacing each observation $\mathbf{x}_i$ with $\mathbf{x}_i - \boldsymbol{\theta}_0$. Many nonparametric methods have been developed as a reaction to the normal-theory based approach of Hotelling's T^2 test. Tests based on marginal signs and ranks (Puri and Sen, 1971), spatial signs and ranks (Oja, 2010), and Oja signs and ranks (Oja, 1999), the optimal signed-rank score tests by Hallin & Paindaveine (Hallin and Paindaveine, 2002a, 2002b), and tests using marginal signs and ranks in the symmetric independent component (IC) model (Nordhausen, Oja, and Paindaveine, 2009) are presented. These tests act as nonparametric and robust competitors to the standard multivariate location inference methods. The parametric Hotelling's T^2 test (Hotelling, 1931) serves as a reference test. The properties of the tests and methods to attain affine invariant test versions from tests not inherently affine invariant are discussed.

A general strategy in the multivariate data analysis methods to be presented is to replace the original observations $\mathbf{x}_i$ with some scores $\mathbf{T}_i = \mathbf{T}(\mathbf{x}_i)$, $i = 1, \ldots, n$. The statistical tests are then based on the new data matrix $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_n)'$. We first go through the parametric Hotelling's T^2 test and then move on to different nonparametric tests all based on different generalized notions of univariate signs and ranks. The tests are put into practice in Chapter 6.

5.2 Hotelling's T^2 test

The classical parametric procedure for the one-sample location problem is Hotelling's T^2 test. Hotelling's T^2 test is any statistical test in which the test statistic follows Hotelling's T^2 distribution. The distribution was developed by Hotelling (1931). Hotelling's T^2 test is a multivariate generalization of the univariate t-test. Let $N_d(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ denote the d-variate normal distribution with

location $\boldsymbol{\theta}$ and covariance $\boldsymbol{\Sigma}$. Furthermore, let $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim N_d(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ be n independent d-variate observations following the multivariate normal distribution. The test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is

$$T^2 = n\bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}},$$

where n is the sample size, $\bar{\mathbf{X}}$ is the vector of sample means, and $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix $\mathbf{S}$. T^2-distribution is proportional to the F-distribution:

$$T^2 \sim \frac{(n-1)d}{n-d} F(d, n-d),$$

where $F(d, n-d)$ is the F-distribution with parameters d and $n-d$. The null hypothesis is rejected at level α if the observed

$$T^2 > \frac{(n-1)d}{n-d} F_\alpha(d, n-d),$$

where $F_\alpha(d, n-d)$ is the α-upper quantile of the F-distribution with parameters d and $n-d$. The test estimate is the sample mean vector. Hotelling's test statistic is affine invariant. Therefore, it has the property

$$T^2(\mathbf{AX}) = T^2(\mathbf{X})$$

for all $d \times d$ full-rank matrices $\mathbf{A}$. The mean serves as an affine equivariant compantion estimator to the test.

If the F-distribution is used as the test statistic, it is assumed that the data are normally distributed. However, for large n, T^2-distribution is approximately χ^2-distributed with d degrees of freedom. If the χ^2-approximation is used, then the normal assumption can be relaxed to the existence of second moments. Thus, Hotelling's T^2 test is an asymptotically distribution-free test.

Hotelling's T^2 test and the sample mean are optimal in the presence of underlying normality. However, they are extremely sensitive to outlying observations and inefficient for heavy-tailed distributions. For these reasons, the goal in research has been to find methods for the one-sample location problem that are valid under much weaker conditions than Hotelling's T^2 test. We now turn to discuss some of these.

5.3 Tests based on marginal signs and ranks

First, we consider multivariate analysis methods based on marginal signs and ranks. Marginal signs and ranks were presented in Section 4.2. The book by Puri and Sen (1971) gives a comprehensive presentation on the subject. Marginal sign- and rank-based approach is based on the criterion functions

$$\mathbf{AVE}_i\{|x_{i1}| + \cdots + |x_{id}|\} \quad \text{and} \quad \mathbf{AVE}_{i,j}\{|x_{i1} - x_{j1}| + \cdots + |x_{id} - x_{jd}|\}.$$

The resulting estimates are the vectors of marginal medians and the marginal Hodges-Lehmann estimates. These are not true multivariate location statistics since they are not affine equivariant. Let $\mathbf{K}(\mathbf{u}) = (K_1(u_1), \ldots, K_d(u_d))'$ be a d-variate vector of score functions. $K_1, \ldots, K_d$ are required to (i) be continuous, (ii) satisfy $\int_0^1 (K_r(\mathbf{u}))^{2+\delta} du < \infty$ for some $\delta > 0$, and (iii) be expressable as the difference of two monotone increasing functions. The test statistic (K-score version) for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is given by

$$Q_{\mathbf{K}} = n \cdot \mathbf{T}_1' \mathbf{B}_1^{-1} \mathbf{T}_1,$$

where $\mathbf{T}_1 = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_1(\mathbf{x}_i) \odot \mathbf{K}\left(\frac{\mathbf{R}_1(\mathbf{x}_i)}{n+1}\right)$ ($\odot$ denotes the Hadamard product, that is, the entrywise product) is the average of the signed-ranks and $\mathbf{B}_1 = \{b_{ij}\}$ is the sample covariance matrix of the signed-ranks with elements

$$b_{ij} = \frac{1}{n} \sum_{k=1}^n S(x_{ki}) S(x_{kj}) K_i\left(\frac{R(x_{ki})}{n+1}\right) K_j\left(\frac{R(x_{kj})}{n+1}\right).$$

Under the null hypothesis, $Q_{\mathbf{K}}$ is asymptotically $\chi^2(d)$-distributed. All d score functions are usually chosen to be the same. The marginal sign test statistic Q_{MS}^2 is obtained with the score function $K_i(u) = 1,\ i = 1, \ldots, d$. Therefore,

$$Q_{MS}^2 = n \left(\frac{1}{n} \sum_{i=1}^n \mathbf{S}_1(\mathbf{x}_i)\right)' \mathbf{B}_1^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{S}_1(\mathbf{x}_i)\right),$$

where the elements of $\mathbf{B}_1$ are

$$b_{ij} = \frac{1}{n} \sum_{k=1}^n S(x_{ki}) S(x_{kj}).$$

The marginal Wilcoxon signed-rank test statistic Q_{MR}^2 is in turn obtained with the score function $K_i(u) = u,\ i = 1, \ldots, d$. Thus,

$$Q_{MR}^2 = n \left(\frac{1}{n(n+1)} \sum_{i=1}^n [\mathbf{S}_1(\mathbf{x}_i) \odot \mathbf{R}_1(\mathbf{x}_i)]\right)' \mathbf{B}_1^{-1} \left(\frac{1}{n(n+1)} \sum_{i=1}^n [\mathbf{S}_1(\mathbf{x}_i) \odot \mathbf{R}_1(\mathbf{x}_i)]\right),$$

where the elements of $\mathbf{B}_1$ are

$$b_{ij} = \frac{1}{n(n+1)^2} \sum_{k=1}^n S(x_{ki}) S(x_{kj}) R(x_{ki}) R(x_{kj}).$$

Unfortunately, tests based on marginal signs and ranks are not invariant under affine transformations of the data. This lack of affine-invariance is one of the main motivations for the other approaches to generalize the univariate sign and rank methods. However, affine invariant versions of marginal sign and signed-rank tests can be obtained by using specific data transformation techniques. One technique is the so-called transformation-retransformation technique introduced by Chakraborty and Chaudhuri (1996) and Chakraborty, Chaudhuri and

Oja (1998). In this technique, the data is first linearly transformed to a new invariant coordinate system and then the marginal test or estimate is constructed on the transformed coordinates. Finally, estimates can then be retransformed to the original coordinate system. The technique applied in this thesis is the invariant coordinate selection (ICS) technique described in Nordhausen et al. (2006) and Nordhausen et al. (2008). Invariant versions of marginal sign and marginal signed-rank tests are obtained if the multivariate variables are first transformed to invariant coordinates, and the univariate sign and rank test is then applied to these transformed variables. In the one-sample location problem, the ICS transformation is based upon the use of two different scatter matrices. It is required that the two scatter matrices are scatter matrices with respect to the origin and that they are invariant under permutations of the data points. Hence, for $k = 1, 2$, it is required that

$$\mathbf{\Sigma}_k(\mathbf{AXPJ}) = \mathbf{A}\mathbf{\Sigma}_k(\mathbf{X})\mathbf{A}'$$

for any nonsingular matrix $\mathbf{A}$, permutation matrix $\mathbf{P}$, and sign-change matrix $\mathbf{J}$. An invariant coordinate system can be found as follows: with two scatter matrices $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_1(\mathbf{X})$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2(\mathbf{X})$, define a $d \times d$ transformation matrix $\mathbf{B} = \mathbf{B}(\mathbf{X})$ and a diagonal matrix $\mathbf{D} = \mathbf{D}(\mathbf{X})$ by

$$\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\mathbf{B}' = \mathbf{B}'\mathbf{D}.$$

Now, the transformation $\mathbf{X} \to \mathbf{Z} = \mathbf{B}(\mathbf{X})\mathbf{X}$ yields an invariant coordinate system in the sense that

$$\mathbf{B}(\mathbf{AX})(\mathbf{AX}) = \mathbf{JB}(\mathbf{X})\mathbf{X}$$

for some $n \times n$ sign-change matrix $\mathbf{J}$.

There are a lot of possible choices of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. So far, there are no exact guidelines about the optimal choice. The choice may depend on the application at hand. However, for most data sets, the different choices yield only minor differences. Also, the two scatter matrices are interchangeable. In general, ICS is easier to apply than the transformation retransformation technique (Nordhausen et al., 2008).

5.4 Affine invariance of spatial sign and signed-rank tests

Next, we turn to multivariate inference methods based on spatial signs and ranks. Spatial signs and ranks were presented in Section 4.3. The book by Oja (2010) gives a thorough review of spatial signs and ranks and related test procedures. The spatial sign- and rank-based tests improve over the marginal sign and rank tests in terms of efficiency but not in terms of affine invariance.

The spatial sign test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is

$$\mathbf{T}_2 = \mathbf{T}_2(\mathbf{X}) = \mathbf{AVE}\{\mathbf{S}_2(\mathbf{x}_i)\},$$

where $\mathbf{S}_2(\cdot)$ denotes the spatial sign function. Under the null hypothesis, its quadratic form is

$$Q_{SS}^2 = n \cdot \mathbf{T}_2' \mathbf{B}_2^{-1} \mathbf{T}_2 \xrightarrow{D} \chi^2(d),$$

where the covariance matrix

$$\mathbf{B}_2 = \mathbf{AVE}\{\mathbf{S}_2(\mathbf{x}_i)(\mathbf{S}_2(\mathbf{x}_i))'\}.$$

The estimate corresponding to the spatial sign test is the so-called spatial median. The (sample) spatial median is the point $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ that minimizes

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\theta}\|$$

– or, equivalently, is the solution to the equation $\mathbf{R}_2(\mathbf{x}) = \mathbf{0}$. The spatial median is location-change equivariant and orthogonal equivariant, but not affine equivariant which is a huge drawback. Magyar and Tyler (2011) show that the spatial median has its highest asymptotic efficiency at spherically symmetric distributions. It is also shown that one can construct an affine equivariant version of the spatial median which is asymptotically more efficient than the regular spatial median. For $d = 1$, the spatial median reduces to the standard univariate median.

The spatial signed-rank test is the multivariate analogue of the univariate Wilcoxon signed-rank test. The spatial signed-rank test statistic for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is the average of spatial signed-ranks, that is,

$$\hat{\mathbf{T}}_2 = \hat{\mathbf{T}}_2(\mathbf{X}) = \mathbf{AVE}\{\mathbf{Q}_2(\mathbf{x}_i)\},$$

where $\mathbf{Q}_2(\cdot)$ denotes the spatial signed-rank function. Under the null hypothesis, its quadratic form is

$$Q_{SR}^2 = n \cdot \hat{\mathbf{T}}_2' \hat{\mathbf{B}}_2^{-1} \hat{\mathbf{T}}_2 \xrightarrow{D} \chi^2(d),$$

where the covariance matrix

$$\hat{\mathbf{B}}_2 = \mathbf{AVE}\{\mathbf{Q}_2(\mathbf{x}_i)(\mathbf{Q}_2(\mathbf{x}_i))'\}.$$

The estimate corresponding to the spatial signed-rank test is the (sample) spatial Hodges-Lehmann estimate which is a multivariate extension of the one-sample Hodges-Lehmann estimate. The spatial Hodges-Lehmann estimate is the point $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ that minimizes

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{x}_i + \mathbf{x}_j - 2\boldsymbol{\theta}\|.$$

Tests based on spatial signs and ranks only require the null distribution of $\mathbf{X}$ to be centrally symmetric about $\boldsymbol{\theta}$. Unfortunately, spatial signs, ranks,

and signed-ranks are only orthogonal equivariant, not affine equivariant, and the corresponding tests are not affine invariant. However, inner standardization can be used to construct affine invariant test versions. The following presentation of affine invariant spatial sign and signed-rank test versions is derived from Oja (2010). An affine invariant test version is obtained by pretransforming the data with a suitable scatter matrix. It has been shown that for any scatter matrix with respect to the origin $\mathbf{\Sigma}$, $Q^2(\mathbf{X}\mathbf{\Sigma}^{-1/2})$ is affine invariant. Inner standardization can be achieved with Tyler's transformation matrix $\mathbf{\Sigma}^{-1/2}$ (Tyler, 1987). By definition, Tyler's transformation matrix $\mathbf{\Sigma}^{-1/2}$ is the matrix that makes the spatial sign and signed-rank covariance matrices proportional to the identity matrix. Thus, one needs to find a transformation matrix $\mathbf{\Sigma}^{-1/2}$ such that if $\hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i),\ i = 1, \ldots, n$, for some d-vector valued score function $\mathbf{T}(\cdot)$ and

$$\hat{\mathbf{T}} = (\hat{\mathbf{T}}_1, \ldots, \hat{\mathbf{T}}_n)',$$

then

$$\hat{\mathbf{T}}'\hat{\mathbf{T}} \propto \mathbf{I}_d.$$

The test statistic is obtained as follows:

$$\mathbf{X} \ \rightarrow \ \hat{\mathbf{T}} \ \rightarrow \ Q^2(\mathbf{X}\mathbf{\Sigma}^{-1/2}) = \mathbf{1}_n'\hat{\mathbf{T}}(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{1}_n.$$

$Q^2(\mathbf{X}\mathbf{\Sigma}^{-1/2})$ is now an affine invariant version of the test statistic.

The spatial signs of the Tyler transformed observations, $\hat{\mathbf{S}}_2(\mathbf{x}_i) = \mathbf{S}_2(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i)$, $i = 1, \ldots, n$, are called standardized spatial signs. Now $\hat{\mathbf{S}}_2 = (\hat{\mathbf{S}}_2(\mathbf{x}_1), \ldots, \hat{\mathbf{S}}_2(\mathbf{x}_n))'$ is the matrix of observed standardized spatial signs and $\hat{\mathbf{S}}_2'\hat{\mathbf{S}}_2 \propto \mathbf{I}_d$. The multivariate spatial sign test based on standardized spatial signs is the spatial sign test with inner standardization. The test rejects $H_0 : \boldsymbol{\theta} = \mathbf{0}$ for large values of

$$Q^2_{SS}(\mathbf{X}\mathbf{\Sigma}^{-1/2}) = \mathbf{1}_n'\hat{\mathbf{S}}_2(\hat{\mathbf{S}}_2'\hat{\mathbf{S}}_2)^{-1}\hat{\mathbf{S}}_2'\mathbf{1}_n = \frac{d}{n} \cdot \|\mathbf{1}_n'\hat{\mathbf{S}}_2\|^2 = nd \cdot \|\mathbf{AVE}\{\hat{\mathbf{S}}_2(\mathbf{x}_i)\}\|^2.$$

The test statistic is affine invariant and, under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$,

$$Q^2_{SS}(\mathbf{X}\mathbf{\Sigma}^{-1/2}) \xrightarrow{D} \chi^2(d).$$

In the elliptic model, $Q^2_{SS}(\mathbf{X}\mathbf{\Sigma}^{-1/2})$ is strictly distribution-free and asymptotically equivalent with the sign test using the affine equivariant Oja signs (Section 5.5).

Analogously, the spatial signed-ranks of the Tyler transformed observations, $\hat{\mathbf{Q}}_2(\mathbf{x}_i) = \mathbf{Q}_2(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i),\ i = 1, \ldots, n$, are called standardized spatial signed-ranks. Now $\hat{\mathbf{Q}}_2 = (\hat{\mathbf{Q}}_2(\mathbf{x}_1), \ldots, \hat{\mathbf{Q}}_2(\mathbf{x}_n))'$ is the matrix of observed standardized spatial signed-ranks and $\hat{\mathbf{Q}}_2'\hat{\mathbf{Q}}_2 \propto \mathbf{I}_d$. The multivariate spatial signed-rank test based on standardized spatial signed-ranks is the spatial signed-rank test with inner standardization. The test rejects $H_0 : \boldsymbol{\theta} = \mathbf{0}$ for large values of

$$Q^2_{SR}(\mathbf{X}\mathbf{\Sigma}^{-1/2}) = \mathbf{1}_n'\hat{\mathbf{Q}}_2(\hat{\mathbf{Q}}_2'\hat{\mathbf{Q}}_2)^{-1}\hat{\mathbf{Q}}_2'\mathbf{1}_n = nd \cdot \frac{\|\mathbf{AVE}\{\hat{\mathbf{Q}}_2(\mathbf{x}_i)\}\|^2}{\mathbf{AVE}\{\|\hat{\mathbf{Q}}_2(\mathbf{x}_i)\|^2\}}.$$

The test statistic is affine invariant and, under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$,

$$Q^2_{SR}(\mathbf{X}\boldsymbol{\Sigma}^{-1/2}) \xrightarrow{D} \chi^2(d).$$

In short, the spatial sign test and the spatial signed-rank test with inner standardizations are affine invariant and the estimates are affine equivariant. As discussed earlier, Magyar and Tyler (2011) showed that one can construct an affine equivariant version of the spatial median which is asymptotically more efficient than the regular spatial median. One can expect that that same holds for the affine invariant versions of the spatial sign and signed-rank tests. Therefore, it can be concluded that using affine invariant spatial sign and signed-rank test versions is always preferable to using non-affine invariant ones.

5.5 Tests based on Oja signs and ranks

Multivariate location test procedures based on Oja signs and signed-ranks are considered next. Oja signs and ranks were presented in Section 4.4. An overview of Oja signs and ranks and corresponding test procedures can be found in Oja (1999). Unlike in the case of marginal and spatial signs and ranks, the test/estimation procedures are now fully affine invariant/equivariant which is a big advantage. These tests only require the null distribution of $\mathbf{X}$ to be centrally symmetric about $\boldsymbol{\theta}$. However, the Oja methods are computationally expansive. One reason is that the number of hyperplanes needed to compute the test statistics increases quickly as the sample size and the number of dimensions increase.

The Oja sign test statistic for testing the hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is the sum of the Oja signs of the observations − or, equivalently, the centered Oja rank of the null hypothesis point, that is,

$$\mathbf{T}_3 = \mathbf{T}_3(\mathbf{X}) = \sum_{i=1}^{n} \mathbf{S}_3(\mathbf{x}_i) \qquad (\sim \mathbf{R}_3(\mathbf{0})),$$

where $\mathbf{S}_3(\cdot)$ denotes the Oja sign function and $\mathbf{R}_3(\cdot)$ denotes the Oja rank function. The quadratic form is

$$Q^2_{OS} = n^{-1}\mathbf{T}_3'\mathbf{B}_3^{-1}\mathbf{T}_3,$$

where the covariance matrix

$$\mathbf{B}_3 = \mathbf{AVE}\{\mathbf{S}_3(\mathbf{x}_i)(\mathbf{S}_3(\mathbf{x}_i))'\}.$$

Analogously, the one-sample Oja signed-rank test statistic is

$$\hat{\mathbf{T}}_3 = \hat{\mathbf{T}}_3(\mathbf{X}) = \sum_{i=1}^{n} \mathbf{Q}_3(\mathbf{x}_i),$$

where $\mathbf{Q}_3(\cdot)$ denotes the Oja signed-rank function. The quadratic form is

$$Q^2_{OR} = n^{-1}\hat{\mathbf{T}}_3'\hat{\mathbf{B}}_3^{-1}\hat{\mathbf{T}}_3,$$

where the covariance matrix

$$\hat{\mathbf{B}}_3 = \mathbf{AVE}\{\mathbf{Q}_3(\mathbf{x}_i)(\mathbf{Q}_3(\mathbf{x}_i))'\}.$$

The estimate corresponding to the Oja tests is the Oja median (Oja, 1983). Oja median is the point $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ that minimizes the criterion function $D_2(\mathbf{x})$ (Section 4.4) or is the solution of $\mathbf{R}_3(\mathbf{x}) = \mathbf{0}$. It is located among the intersection points of the observation hyperplanes. The computation of the estimate is highly intensive. For a discussion of the different algorithms to compute the Oja median, see Ronkainen, Oja, and Orponen (2003). For $d = 1$, the Oja median reduces to the standard univariate median.

In the ellipitic case, the sign test using the affine equivariant Oja signs is asymptotically equivalent to the invariant version of the spatial sign test using $Q^2_{SS}(\mathbf{X}\boldsymbol{\Sigma}^{-1/2})$. The latter is computationally much more convenient. However, at the elliptic model, their efficiency may be poor when compared with Hallin and Paindaveine tests presented next (Oja, 2010). Still, Oja methods remain valid under the weaker central symmetry assumption.

5.6 The optimal signed-rank scores tests by Hallin and Paindaveine

Another generalization of the univariate sign and rank test procedures are the optimal signed-rank score tests proposed by Hallin and Paindaveine (2002a, 2002b). These tests combine the ranks of pseudo-Mahalanobis distances between the data points and their center $\boldsymbol{\theta}$ either with Randles' interdirections (Randles, 1989) or with the so-called Tyler's angles. Tests based on interdirections are described in Hallin and Paindaveine (2002a) and those based on Tyler's angles in Hallin and Paindaveine (2002b). Both test versions are asymptotically equivalent. The assumption in both tests is that the data comes from an elliptic distribution. These tests have the desired property of affine invariance. Thus, for example, the p-value does not depend on the chosen coordinate system.

Interdirections could be described as a fully hyperplane-based class of procedures for the one-sample location problem. Interdirections measure the angular distance between two observation vectors relative to the rest of the data. The interdirection $c_{ij} \in \mathbb{N}$ associated with the pair $(\mathbf{x}_i, \mathbf{x}_j)$ is defined as the number of hyperplanes in $\mathbb{R}^d$ passing through the origin and $d-1$ out the $n-2$ points $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_n$ that separate $\mathbf{x}_i$ and $\mathbf{x}_j$, $0 \leq c_{ij} \leq \binom{n-2}{d-1}$, $i, j = 1, \ldots, n$. Let $p_{ij} = c_{ij}/\binom{n}{d-1}$, $0 \leq p_{ij} \leq 1$, be the proportion of such (data-based) hyperplanes that pass through $\mathbf{x}_i$ and $\mathbf{x}_j$. Hallin and Paindaveine (2002b) present a class of test statistics (K-score version) based

on interdirections and pseudo-Mahalanobis ranks (equation (2.2) in Hallin and Paindaveine, 2002b). The interdirection-based sign test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is obtained by letting the score function $K(u) = 1$. Thus, we obtain

$$Q^2_{\text{HPIS}} = \frac{d}{n} \sum_{i,j=1}^{n} \cos(\pi p_{ij}).$$

The null hypothesis is rejected at level α if $Q^2_{\text{HPIS}} > \chi^2_\alpha(d)$. For $d = 1$, the test statistic reduces to the regular univariate sign test statistic.

Interdirections together with a ranking of the magnitudes of the observations can be used to generalize the univariate signed-rank tests. The signed-rank test proposed by Hallin and Paindaveine (2002a) is based on the ranks of pseudo-Mahalanobis distances between the observations and their center $\boldsymbol{\theta}$. Let $d_i = d_i(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \left((\mathbf{x}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\theta})\right)^{1/2}$, $i = 1, \ldots, n$, denote the distances between $\mathbf{x}_i$ and $\boldsymbol{\theta}$ in the metric associated with $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is Tyler's scatter matrix. Furthermore, let R_i denote the rank of d_i among all the distances $d_1, \ldots, d_n$. Write $\hat{R}_i$ and $\hat{d}_i$ for the data-based R_i and d_i. $\hat{R}_i$ is the pseudo-Mahalanobis rank of $\mathbf{x}_i$. By choosing the score function $K(u) = au$, $u \in]0, 1[$, $a > 0$, we obtain the interdirection-based signed-rank test statistic. The null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is rejected at level α if

$$Q^2_{\text{HPIR}} = \frac{3d}{n(n+1)^2} \sum_{i,j=1}^{n} \hat{R}_i \hat{R}_j \cos(\pi p_{ij}) > \chi^2_\alpha(d).$$

These tests do not require any moment assumptions and relative to the Hotelling's T^2 test, they offer broader validity and better robustness features, that is, better resistance to violations of the assumptions.(Hallin and Paindaveine, 2002a).

An alternative to the interdirection-based testing procedures are those based on Tyler's angles (Hallin and Paindaveine, 2002b). Tyler's angles are the angles between the observations standardized via Tyler's estimator of scatter. The resulting tests are called angle-based tests and they are valid under the same class of densities as the interdirection-based tests. Let $\hat{\mathbf{S}}_2(\mathbf{x}_i) = \mathbf{S}_2(\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i)$, $i = 1, \ldots, n$, denote the spatial signs of the standardized observations $\mathbf{x}_i$, where $\boldsymbol{\Sigma}$ is Tyler's scatter matrix . The equation (3.1) in Hallin and Paindaveine, 2002b is the "Tyler" analog of the test statistic (2.2) in Hallin and Paindaveine, 2002b. Our Tyler sign test and signed-rank test statistics are obtained with the same score functions as in the case of interdirection-based tests. Thus, the Tyler sign test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is

$$Q^2_{HPTS} = \frac{d}{n} \sum_{i,j=1}^{n} (\hat{\mathbf{S}}_2(\mathbf{x}_i))'\hat{\mathbf{S}}_2(\mathbf{x}_j),$$

and the Tyler signed-rank test statistic is

$$Q^2_{HPTR} = \frac{3d}{n(n+1)^2} \sum_{i,j=1}^{n} \hat{R}_i \hat{R}_j (\hat{\mathbf{S}}_2(\mathbf{x}_i))'\hat{\mathbf{S}}_2(\mathbf{x}_j).$$

Again, the null hypothesis is rejected at level α if the value of the test statistic exceeds $\chi^2_\alpha(d)$. A little manipulation shows that the Tyler sign test statistic Q^2_{HPTS} is equivalent to the invariant version of the spatial sign test using $Q^2_{SS}(\mathbf{X}\boldsymbol{\Sigma}^{-1/2})$.

Angle-based procedures and interdirection-based procedures share the same invariance properties and asymptotic efficiencies. However, angle-based procedures are computationally preferable to interdirections because they avoid the computation of interdirections. Calculating interdirections is computationally heavy, especially in higher dimensions. For this reason, optimal signed-rank score tests based on Tyler's angles are used in the simulation study in Chapter 6.

5.7 Tests using marginal signs and ranks in the symmetric IC model

Lastly, we apply marginal sign- and rank-based tests on data that are assumed to follow the symmetric independent component (IC) model. Using marginal signs and ranks in the symmetric independent component model is presented in Nordhausen, Oja, and Paindaveine (2009). The symmetric independent component model is a special case of the location-scatter model and an extension of the multivariate normal model. In the location-scatter model, the d-variate random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are generated by

$$\mathbf{X}_i = \boldsymbol{\Lambda}\mathbf{Z}_i + \boldsymbol{\theta}, \qquad i = 1, \ldots, n,$$

where the vectors $\mathbf{Z}_i$ are standardized in some way. The full-rank $d \times d$ matrix $\boldsymbol{\Lambda}$ is called the mixing matrix and the parameter $\boldsymbol{\theta}$ is the location center. Different standardizations of the $\mathbf{Z}_i$'s lead to different location-scatter models. In the multinormal model, it is assumed that $\mathbf{Z}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$. In the elliptic model, $\mathbf{Z}_i$ has a spherical distribution around the origin with $\text{Med}[\|\mathbf{Z}_i\|^2] = \chi^2_{0.5}(d)$, where Med$[\cdot]$ denotes the population median and $\chi^2_\alpha(d)$ denotes the α-quantile of the $\chi^2(d)$-distribution. In the symmetric independent component model, the components of $\mathbf{Z}_i$ are independent and symmetric (that is, $-Z_{ik} \sim Z_{ik}$) with $\text{Med}[(Z_{ik})^2] = \chi^2_{0.5}(1),\ k = 1, \ldots, d$. The IC model can be formulated in many ways: if the independent components are permuted or multiplied by nonzero scalars, they still remain independent. The problem of estimating $\boldsymbol{\Lambda}$ in this model is known as the independent component analysis (ICA).

The tests discussed are signed-rank tests (with constant and identity score functions) applied to the residuals $\mathbf{Z}(\boldsymbol{\Lambda}_i) = \hat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}_i,\ i = 1, \ldots, n$, where $\hat{\boldsymbol{\Lambda}}$ is a suitable estimate of the mixing matrix $\boldsymbol{\Lambda}$. The first problem is to find such estimate $\hat{\boldsymbol{\Lambda}}$. In the unrealistic case that $\boldsymbol{\Lambda}$ is known, there is no problem. If $\boldsymbol{\Lambda}$ is not known, then any esimate $\hat{\boldsymbol{\Lambda}}$ that is root-n consistent and invariant under individual sign changes of the observations can be used (Nordhausen et al., 2009). The mixing matrix $\boldsymbol{\Lambda}$ can be estimated by using two different scatter

matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. When an appropriate estimate $\hat{\mathbf{\Lambda}}$ is found, the marginal signs and ranks are computed for each component. These will be denoted by $\mathbf{S}_1(\mathbf{\Lambda}_i) = (S(\mathbf{\Lambda}_{i1}), \ldots, S(\mathbf{\Lambda}_{id}))'$ and $\mathbf{R}_1(\mathbf{\Lambda}_i) = (R(\mathbf{\Lambda}_{i1}), \ldots, R(\mathbf{\Lambda}_{id}))'$, where $R(\mathbf{\Lambda}_{ij})$ denotes the marginal rank of $|\mathbf{Z}(\mathbf{\Lambda}_{ij})|$ among all $|\mathbf{Z}(\mathbf{\Lambda}_{1j})|, \ldots, |\mathbf{Z}(\mathbf{\Lambda}_{nj})|$. Next, an appropriate score function for each component is chosen. Finally, marginal scores are combined to form a test statistic. The dependence between the components is in the estimated covariance matrix $\mathbf{B}_1$. Naturally, if the components are indendent, the estimated covariance matrix $\mathbf{B}_1$ is converging to a diagonal matrix. Thus, $\mathbf{B}_1$ can be replaced by its probability limit

$$\mathbf{B}_1 = \mathrm{diag}(\mathrm{E}[(K_1(U))^2], \ldots, \mathrm{E}[(K_d(U))^2]),$$

where U is uniformly distributed over $(0, 1)$. The test statistic (K-score version) proposed by Nordhausen et al. (2009) is

$$Q_{\mathbf{K}}(\mathbf{\Lambda}) = (\mathbf{T}_1(\mathbf{\Lambda}))'\mathbf{B}_1^{-1}\mathbf{T}_1(\mathbf{\Lambda}),$$

where $\mathbf{T}_1(\mathbf{\Lambda}) = n^{-1/2}\sum_{i=1}^n \mathbf{T}_1(\mathbf{\Lambda}_i) = n^{-1/2}\sum_{i=1}^n \left[\mathbf{S}_1(\mathbf{\Lambda}_i) \odot \mathbf{K}\left(\frac{\mathbf{R}_1(\mathbf{\Lambda}_i)}{n+1}\right)\right]$ and $\mathbf{B}_1 = \mathrm{diag}(\mathrm{E}[(K_1(U))^2], \ldots, \mathrm{E}[(K_d(U))^2])$. Under H_0, $Q_{\mathbf{K}}(\mathbf{\Lambda})$ is asymptotically χ^2-distributed with d degrees of freedom. Our sign test statistic is obtained with the score function $K_r(u_r) = 1$ for all $r = 1, \ldots, d$, and the signed-rank test statistic is obtained with the identity score function $K_r(u_r) = u_r$ for all $r = 1, \ldots, d$. Let $\mathbf{S}_1(\hat{\mathbf{\Lambda}}_i)$ and $\mathbf{R}_1(\hat{\mathbf{\Lambda}}_i)$, $i = 1, \ldots, n$, denote the empirical signs and ranks. The sign test statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is

$$Q^2_{ICMS} = \frac{1}{n}\sum_{i,j=1}^n (\mathbf{S}_1(\hat{\mathbf{\Lambda}}_i))'\mathbf{S}_1(\hat{\mathbf{\Lambda}}_j) = \frac{1}{n}\sum_{i,j=1}^n \sum_{k=1}^d S(\hat{\mathbf{\Lambda}}_{ik})S(\hat{\mathbf{\Lambda}}_{jk}),$$

and the signed-rank test statistic is

$$Q^2_{ICMR} = \frac{3}{n(n+1)^2}\sum_{i,j=1}^n \sum_{k=1}^d S(\hat{\mathbf{\Lambda}}_{ik})S(\hat{\mathbf{\Lambda}}_{jk})R(\hat{\mathbf{\Lambda}}_{ik})R(\hat{\mathbf{\Lambda}}_{jk}).$$

These tests are affine invariant (given $\hat{\mathbf{\Lambda}}$ is affine equivariant), robust, and they do not require any moment assumptions (Nordhausen et al., 2009).

6 Simulation study

6.1 Introduction

In this chapter, we focus on the practical implementation of our tests and compare their relative performance under different settings. The goal is to provide practical guidelines which test might be most useful in practice. Computations were done using the statistical software package R 3.0.2 (The R Foundation for Statistical Computing). All the methods proposed can easily be applied with R packages ICS, MNM, ICSNP, and OjaNP (see Nordhausen et al. 2013, Nordhausen et al. 2011, Nordhausen et al. 2012, and Fischer et al. 2013).

The simulation data are generated from p-generalized normal distributions and L_p-norm multivariate t-distributions. These distributions and methods to generate random samples from them were presented in Chapter 3. The simulations are based on 1,000 repetitions at level $\alpha = 0.05$ for sample sizes $n = 30, 50, 100, 200$ and dimensions $d = 2, 3, 5$. For p-generalized normal distributions, $p = 0.5, 1, 1.5, 2, 3$ and for L_p-norm multivariate t-distributions, $p = 2, 3$ with degrees of freedom $df = 3$ and $p = 1, 1.5, 2, 3$ with $df = 9$. The limited number of settings regarding L_p-norm multivariate t-distributions is due to the properties of the probability density functions of L_p-norm multivariate t-distributions which require that $(d + df)/2 > d/p$. In addition, in order to be able to compare the different simulation results, the simulation data are standardized so that $\mathrm{cov}(\mathbf{X}) = \mathbf{I}_d$ for any given sample $\mathbf{X}$. This is not a restriction since all the tests involved in the study are affine invariant. The covariance matrices of L_p-norm distributions, provided they exist, have a general form

$$\mathrm{cov}(\mathbf{X}) = \frac{\Gamma(d/p)\Gamma(3/p)}{\Gamma(1/p)\Gamma((d+2)/p)}\mathrm{E}(R^2)\mathbf{I}_d$$

(Gupta and Song 1997). The integral function needed to compute the value of the covariance matrix of L_p-norm multivariate t-distributions (Section 3.2) to standardize the data is convergent only for these value combinations of p and df. The existence of finite second moments also ensures that Hotelling's T^2 test is valid. Furthermore, due to the heaviness of computing Oja signs and signed-ranks, the related tests are performed only for $d = 2$ and for $d = 3$ with $n = 30, 50$ because higher dimensions and sample sizes take too long to compute.

Table 6.1. The shift values Δ.

		$P[\chi^2(d,\delta) > \chi^2_\alpha(d)]$			
d		0.2	0.4	0.6	0.8
2	0.000000	1.315645	1.957628	2.492534	3.103955
3	0.000000	1.447881	2.121483	2.674697	3.301922
5	0.000000	1.633184	2.351308	2.931003	3.581567

Throughout, we apply location testing for $H_0 : \boldsymbol{\theta} = \mathbf{0}$. This is done without loss of generality since location testing about any other fixed value $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ would be obtained by applying the proposed origin-based test to the centered observations $\mathbf{x}_i - \boldsymbol{\theta}_0,\ i = 1, \ldots, n$.

In order to compare the powers of the different tests, the location parameters of the distributions are set to $\boldsymbol{\theta} = \frac{1}{\sqrt{n}}(\Delta, 0, \ldots, 0)'$ and the shift Δ chosen 0 (null case) and in a way that given the dimension d, the power of Hotelling's T^2 test is 0.2, 0.4, 0.6, and 0.8 under normality. This means

$$P[\chi^2(d,\delta) > \chi^2_\alpha(d)] = 0.2,\ 0.4,\ 0.6,\ \text{and } 0.8,$$

where $\chi^2(d,\delta)$ is a random variable having a noncentral χ^2-distribution with d degrees of freedom and a noncentrality parameter $\delta = \Delta^2$, and $\chi^2_\alpha(d)$ is the $1-\alpha$ quantile of $\chi^2(d) = \chi^2(d,0)$. This gives the range of Δ from 0.000 to 3.582. Table 6.1 shows the shift values Δ obtained. Rejection proportions per 1,000 cases estimate the powers of the tests.

Invariant coordinate selection (ICS) is used in marginal sign- and rank-based tests in order to attain affine invariant test versions. ICS uses two scatter matrices to transform the data (Section 5.3). When testing a location parameter, the hypothesis used should be noted in the computation of the scatter matrices (Nordhausen et al., 2008). Since our null hypothesis location is the origin, we use scatter matrices with respect to the origin when creating our ICS. We choose as $\boldsymbol{\Sigma}_1$ the covariance matrix with respect to the origin and as $\boldsymbol{\Sigma}_2$ the scatter matrix based on fourth moments with respect to the origin, that is,

$$\boldsymbol{\Sigma}_1 = E[\mathbf{X}\mathbf{X}'] \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \frac{1}{d+2} E[(\mathbf{X}'\boldsymbol{\Sigma}_1^{-1}\mathbf{X})\mathbf{X}\mathbf{X}'].$$

Both scatter matrices are provided in the R package ICS.

Hotelling's T^2 test is expected to be uniformly the most powerful test in the normal model and also asymptotically valid in the other models because the first two moments exist. Marginal, spatial, and Oja sign and signed-rank tests are valid in all models. The optimal signed-rank score tests by Hallin and Paindaveine (2002b) are in general only valid in the elliptic model ($p = 2$). The spatial sign test with inner standardization and the Oja sign test are asymptotically equivalent in the elliptic model (Oja, 2010). The signed-rank score tests

by Nordhausen et al. (2009) are valid in the symmetric IC model. Samples coming from p-generalized normal distributions follow the symmetric IC model, whereas samples coming from the L_p-norm multivariate t-distributions do not since the p-generalized normal distribution is the only L_p-spherically symmetric distribution with independent marginals (Sinz, Gerwinn, and Bethge, 2009).

6.2 Results

In this section, the most important simulation results are summarized. We first check whether (i) the tests meet the nominal probability level $\alpha = 0.05$ and whether (ii) Hotelling's T^2 test gets everywhere the powers designed in the normal model. We then compare (iii) the powers of the sign and signed-rank tests, (iv) the powers of Hotelling's T^2 test and the powers of the sign and signed-rank tests, and, finally, (v) the powers of all tests. The effects of varying underlying distributions, sample sizes, and dimensions on powers are reported. Due to the large number settings involved in the study, only the most prominent results are highlighted. Complete simulation plots are provided in the Appendix.

6.2.1 Do the tests meet the nominal level $\alpha = 0.05$?

First, it was studied whether the tests attain the advertised level $\alpha = 0.05$. To this end, simulation data were generated under $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and rejection proportions of H_0 were calculated. A glance at the rejection proportions under the null in Figures 1 to 11 in the Appendix shows that all tests appear to satisfy the 5 % probability level constraint in all settings with the exception of Hotelling's T^2 test and the spatial signed-rank test Q^2_{SR} using inner standardization which are consistently biased with sample size $n = 30$ in dimensions $d = 3$, 5 and with sample size $n = 50$ in dimension $d = 5$. Thus, these tests suffer from type I error to some degree which makes the power comparisons harder.

6.2.2 The powers of Hotelling's T^2 test designed in the normal model

Second, it was checked whether Hotelling's T^2 test gets everywhere the powers designed in the normal model. The shift values were chosen so that Hotelling's T^2 test would have powers 0.2, 0.4, 0.6, and 0.8 in the normal model. Rejection proportions of Hotelling's T^2 test in the normal model are reported in Table 6.2. Excluding the settings mentioned above, Hotelling's T^2 test roughly obtains the powers desired.

Table 6.2. The powers of Hotelling's T^2 test in the normal model.

		Δ				
d	n	0	1	2	3	4
2	30	0.059	0.234	0.429	0.613	0.802
	50	0.064	0.210	0.380	0.616	0.819
	100	0.060	0.180	0.401	0.623	0.793
	200	0.050	0.195	0.407	0.596	0.803
3	30	0.101	0.247	0.435	0.670	0.820
	50	0.068	0.237	0.425	0.594	0.810
	100	0.059	0.208	0.411	0.620	0.806
	200	0.055	0.231	0.396	0.623	0.802
5	30	0.142	0.315	0.498	0.703	0.842
	50	0.086	0.266	0.482	0.643	0.816
	100	0.070	0.230	0.457	0.631	0.840
	200	0.060	0.229	0.403	0.604	0.805

6.2.3 Comparison of sign and signed-rank tests

We now turn to comparing the power properties of the sign and signed-rank tests. It should be noted that even though the remaining result sections are divided into parts by the effects of varying (i) underlying distributions, (ii) sample sizes, and (iii) numbers of dimensions on powers, these effects cannot strictly be studied in isolation: the effect of a given variable is more or less dependent on the values of the other variables. Therefore, within each segment, the influence of the other variables is addressed as well.

Effect of p

Figure 6.1 illustrates the behaviour of the sign and signed-ranks tests for data coming from different p-generalized normal distributions ($n = 200$ and $d = 3$). The figure indicates that the sign tests yield higher powers with lower p:s. However, the signed-rank tests gain relatively more power as p increases outperforming the sign tests with higher p:s. Notably, the powers of the non-parametic tests decrease as p increases. Similar results are obtained with other sample sizes and dimensions and when the data comes from L_p-norm multivariate t-distributions.

Effect of n

Figure 6.2 shows the behaviour of the sign and signed-ranks tests for different sample sizes for data coming from the p-generalized normal distribution, $p = 1$

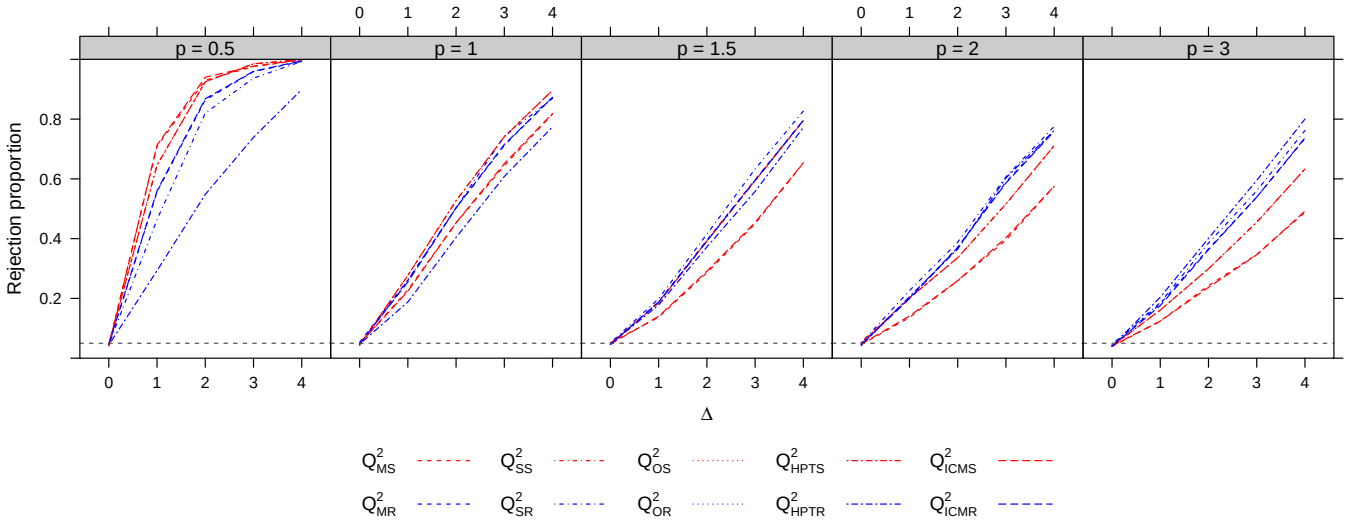

Figure 6.1. Rejection proportions (for $n = 200$ and $d = 3$, based on 1,000 replications) of the sign and signed-rank tests for data coming from different p-generalized normal distributions.

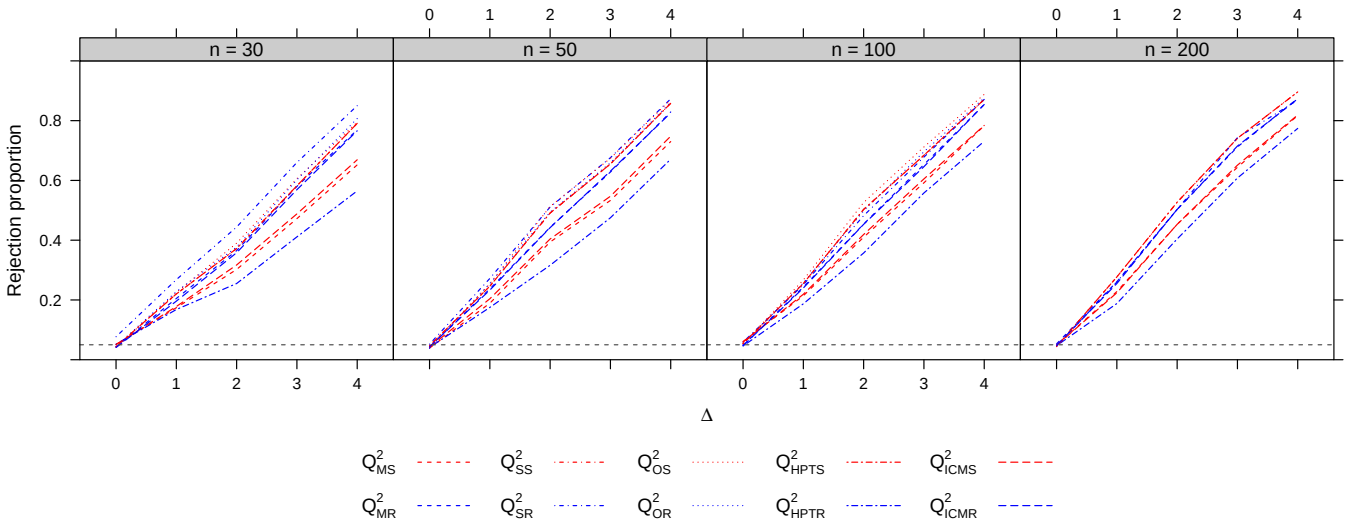

Figure 6.2. Rejection proportions (for $d = 3$, based on 1,000 replications) of the sign and signed-rank tests for different sample sizes for data coming from the p-generalized normal distribution, $p = 1$.

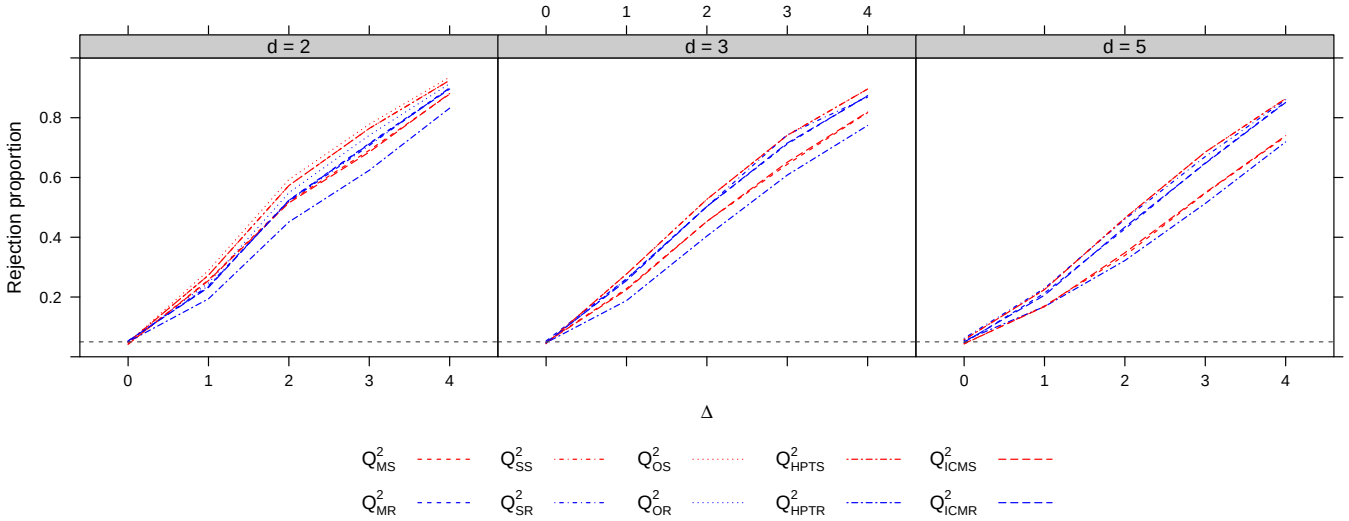

Figure 6.3. Rejection proportions (for $n = 200$, based on 1,000 replications) of the sign and signed-rank tests for different dimensions for data coming the p-generalized normal distribution, $p = 1$.

($d = 3$). The powers increase and get closer to each other as sample size increases. This holds true for distributions with lower p:s, whereas when the data comes from distributions with higher p:s, the powers of the nonparametric tests do not show significant increase with increasing sample size (see the Appendix). In those settings, the powers of the sign and signed-rank tests stay more or less stagnant with increasing n.

Effect of d

Figure 6.3 illustrates the behaviour of the sign and signed-ranks tests for different dimensions for data coming from the p-generalized normal distribution, $p = 1$ ($n = 200$). The figure shows that the power curves diverge a little as the number of dimensions goes up. The powers do not show any significant decrease with increasing d. The same pattern can be seen in others settings, too. The figure also shows that increasing the number of dimensions does not solely benefit either the sign tests or the signed-rank tests over the others.

6.2.4 Comparison of Hotelling's T^2 test and the sign and signed-rank tests

Next, we turn to comparing the power properties of Hotelling's T^2 test relative to the power properties of the sign and signed-ranks tests.

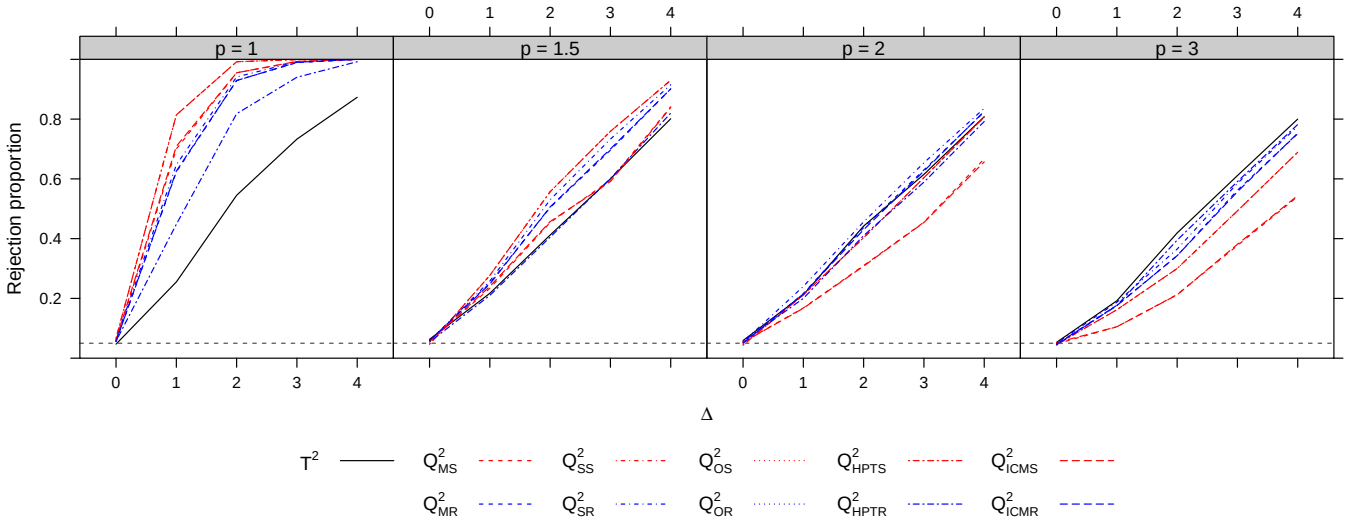

Figure 6.4. Rejection proportions (for $n = 200$ and $d = 3$, based on 1,000 replications) of Hotelling's T^2 test and the sign and signed-rank tests for data coming from different L_p-norm multivariate t-distributions, $df = 9$.

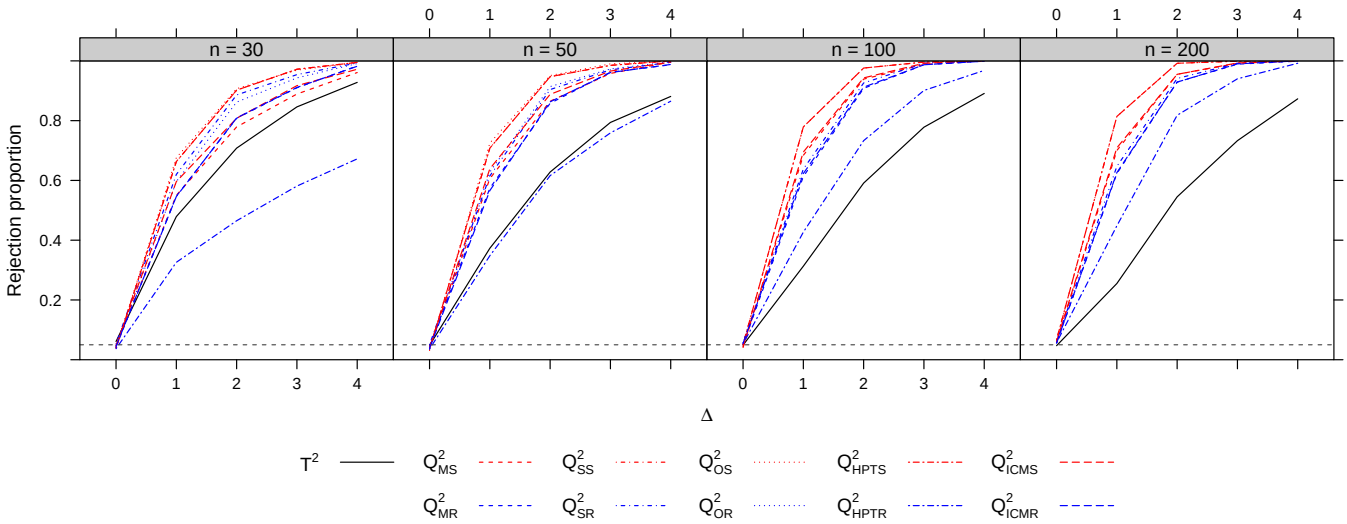

Figure 6.5. Rejection proportions (for $d = 3$, based on 1,000 replications) of Hotelling's T^2 test and the sign and signed-rank tests for different sample sizes for data coming from the L_p-norm multivariate t-distribution, $p = 1$ with $df = 9$.

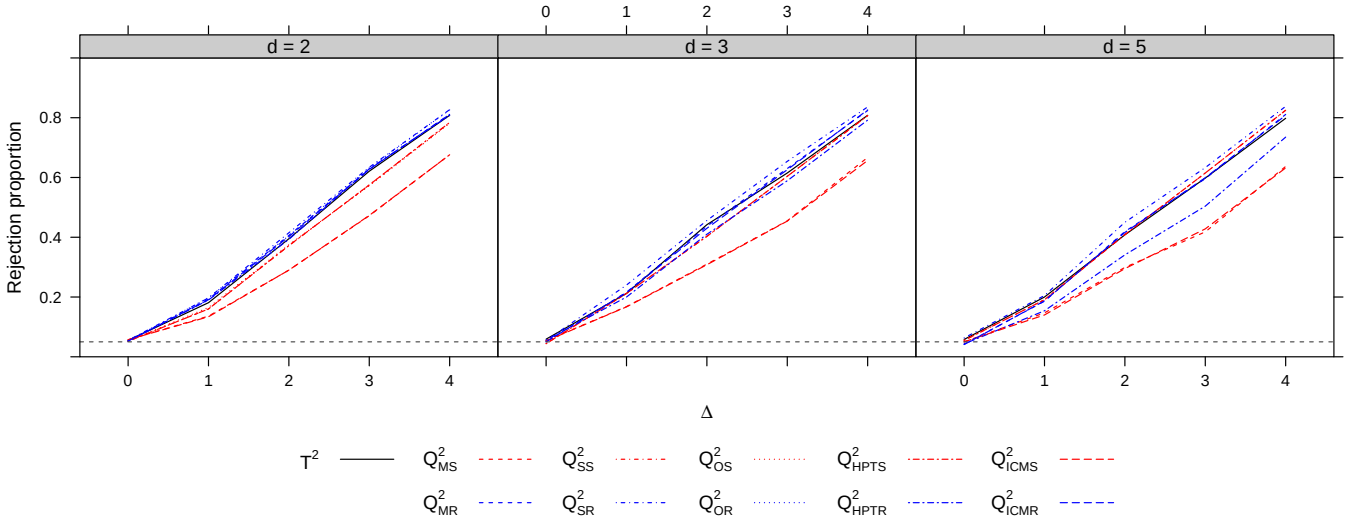

Figure 6.6. Rejection proportions (for $n = 200$, based on 1,000 replications) of Hotelling's T^2 test and the sign and signed-rank tests for different dimensions for data coming from the L_p-norm multivariate t-distribution, $p = 2$ with $df = 9$.

Effect of p

Figure 6.4 shows the behaviour of Hotelling's T^2 test relative to the sign and signed-ranks tests for data coming from different L_p-norm multivariate t-distributions, $df = 9$ ($n = 200$ and $d = 3$). The T^2 test does relatively better as p increases. The powers of the nonparametric tests drop notably with increasing p. Hotelling's T^2 test uniformly outperforms the nonparametric tests with larger p:s. The same pattern holds with other sample sizes and dimensions and when the data comes from p-generalized normal distributions.

Effect of n

Figure 6.5 shows the behaviour of Hotelling's T^2 test relative to the sign and signed-ranks tests for different sample sizes for data coming from the L_p-norm multivariate t-distribution, $p = 1$ with $df = 9$ ($d = 3$). The nonparametric tests gain relatively more power as sample size increases. This is true for data coming from distributions with smaller p:s. In those settings, the T^2 test does relatively better with smaller sample sizes, but as sample size n increases, the nonparametric tests gain relatively more power outperforming the T^2 test. However, with larger p:s, Hotelling's T^2 test dominates uniformly irrespective of sample size, but the nonparametric get closer in power with increasing sample size.

Effect of d

Figure 6.6 illustrates the behaviour of Hotelling's T^2 test relative to the sign and signed-ranks tests for different numbers of dimensions for data coming from the L_p-norm multivariate t-distribution, $p = 2$ with $df = 9$ ($n = 200$). The figure shows that the relative powers of Hotelling's T^2 do not change significantly with varying d. The power curves of the sign and signed-ranks diverge a little with increasing d. In general, with lower p:s, Hotelling's T^2 test does relatively better with increasing d. With higher p:s, the relative positions of the power curves remain largely unaffected by a change in d.

6.2.5 Comparison of all tests

Finally, the power properties of all tests are compared.

Effect of p

Figure 6.7 illustrates the behaviour of all tests for data coming from different p-generalized normal distributions ($n = 200$ and $d = 2$). As discussed before, the powers of the nonparametic tests decrease along with the increase of p, especially those of the sign tests. At the same time, Hotelling's T^2 test gains relative power over the nonparametric tests.

In terms of the sign tests, the marginal sign tests Q^2_{MS} and Q^2_{ICMS} get almost uniformly outperformed by the other sign tests. The marginal sign tests are roughly on par with each other in all settings. With lower p:s, the Oja sign test Q^2_{OS} exhibits highest powers (along with the other sign tests). As p increases, the sign tests gradually lose power relative to the other tests. The marginal sign tests lose most power. With increasing p, the Oja sign test Q^2_{OS} behaves accordingly with the spatial sign test Q^2_{SS} using inner standardization and with the sign test Q^2_{HPTS} based on Tyler's angles. The sign tests get almost uniformly outperformed by Hotelling's T^2 test and the signed-rank tests with higher p:s.

As for the signed-rank tests, the Oja signed-rank test Q^2_{OR} yields highest powers with lower p:s. The signed-rank test Q^2_{HPTR} based on Tyler's angles displays notably lower powers than the other signed-rank tests with lower p:s. The signed-rank tests also lose power with increasing p but not as much as the sign tests. The powers of the signed-rank tests tend to converge as p increases: the tests display very similar powers with higher p:s. Q^2_{HPTR} loses relatively little power with increasing p making it the optimal nonparametic test with $p = 3$. The marginal signed-rank tests Q^2_{MR} and Q^2_{ICMR} do well compared with the other signed-rank tests, but are not the optimal choice in any setting. The order of the power curves of the signed-rank tests almost turns around as p increases. The spatial signed-rank Q^2_{SR} with inner standardization is the best

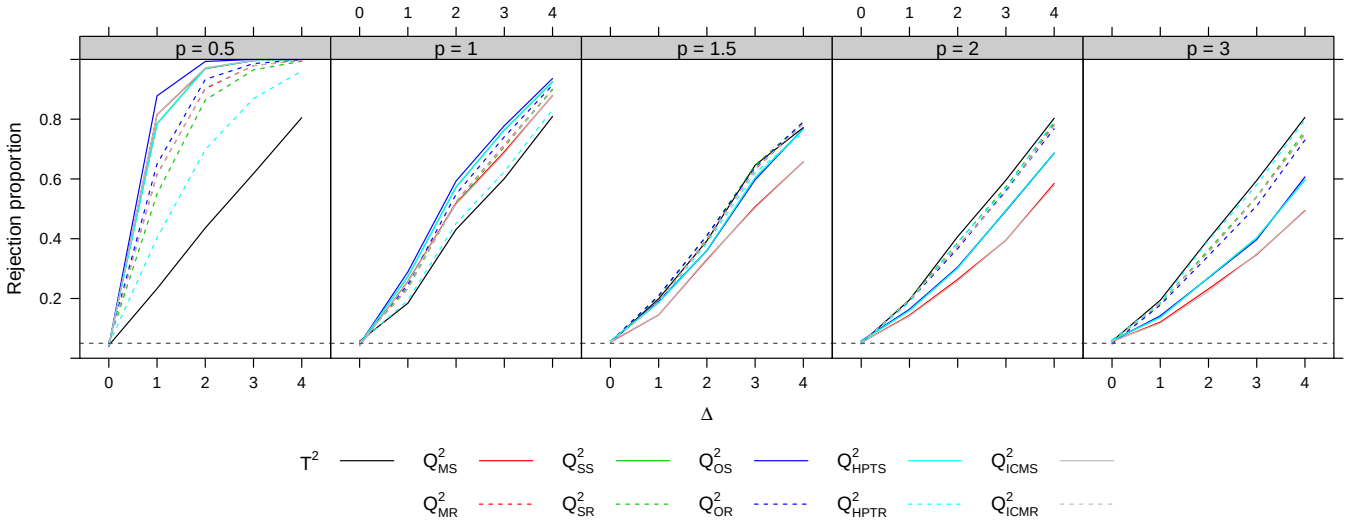

Figure 6.7. Rejection proportions (for $n = 200$ and $d = 2$, based on 1,000 replications) of all tests for data coming from different p-generalized normal distributions.

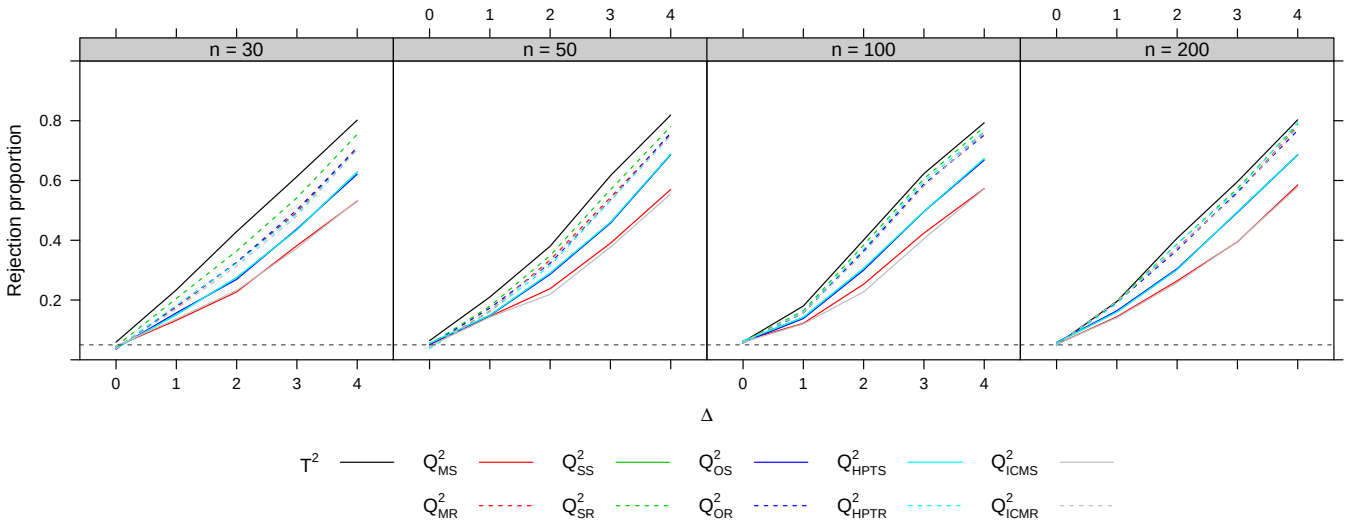

Figure 6.8. Rejection proportions (for $d = 2$, based on 1,000 replications) of all tests for different sample sizes for data coming from the multinormal distribution.

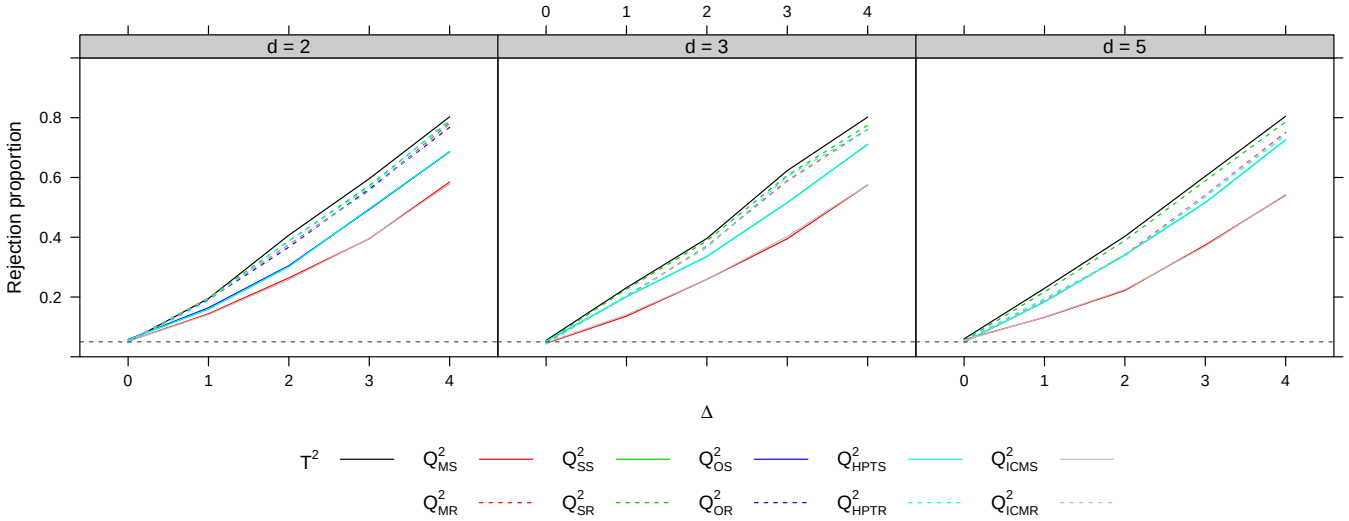

Figure 6.9. Rejection proportions (for $n = 200$, based on 1,000 replications) of all tests for different dimensions for data coming the multinormal distribution.

nonparametric test in the normal case. The signed-rank tests almost uniformly outperform the sign tests with higher p:s.

Hotelling's T^2 loses clearly to the nonparametric tests with lower p:s, but, with increasing p, it gains power over the nonparametric tests outperforming them uniformly with higher p:s. Hotelling's T^2 test is the optimal test in the normal case, as expected.

Simulations based on L_p-norm multivariate t-distributions and other sample sizes and dimensions led to similar results which is why the corresponding plots are not presented here (see the Appendix).

Effect of n

Figure 6.8 shows the behaviour of all tests for different sample sizes for data coming from the multinormal distribution (p-generalized normal distribution, $p = 2$) ($d = 2$). The figure shows that the order of the power curves of the nonparametic tests is not significantly affected by an increase in sample size. In other words, no single sign or signed-rank test benefits notably over other sign or signed-rank tests from an increase in sample size. The order of the power curves is largely determined by the underlying distribution. The power curves tend to converge with increasing sample size. The observations made in the previous section on powers in regard to the effect of sample size hold true: the powers of the nonparametic tests increase relative to Hotelling's T^2 test as sample size increases. However, the nonparametric tests do not exceed Hotelling's T^2 test in power since, with larger p:s, varying sample size has less effect on powers. With lower p:s, the nonparametric tests gain relatively more power than Hotelling's T^2 test as sample size increases outperforming the T^2

test with higher sample sizes. The same pattern holds for other dimensions and samples coming from L_p-norm multivariate t-distributions.

Effect of d

Figure 6.9 illustrates the behaviour of all tests for different numbers of dimensions for data coming from the multinormal distribution ($n = 200$). The marginal sign tests Q^2_{MS} and Q^2_{ICMS} lose some power relative to the other tests with increasing d. The other tests remain relatively unaffected by a change in number of dimensions. Similar results are obtained in other settings. One exception is that Hotelling's T^2 test does relatively better with lower p:s as d increases (also, the signed-rank test Q^2_{HPTR} based on Tyler's angles loses power with lower p:s as d increases). With higher p:s, the relative positions of the power curves remain largely unnaffected by a change in d. The order of the power curves is largely determined by the underlying distribution. Hotelling's T^2 dominates in the normal case irrespective of the number of dimensions.

7 Conclusions

This thesis discusses different tests considered in the literature for the one-sample location problem. A simulation study was conducted to compare the power properties of the tests. The main findings from this study can be summarized as follows: (i) none of the location tests is superior to all others in all settings. (ii) The underlying distribution was found to have the biggest impact on powers. Notably, the powers of the nonparametric tests drop as p increases. Also, the relative order of the power curves of the sign and signed-rank tests changes as p increases: the sign tests display higher powers with lower p:s, but with increasing p, the signed-rank tests outperform the sign tests. At the same time, Hotelling's T^2 test gains power over the nonparametric tests. (iii) The underlying distribution was also found to largely determine how much influence the sample size and the number of dimensions have on powers. In regard to sample size, with lower p:s, Hotelling's T^2 test does relatively better with smaller sample sizes, but with increasing n, the nonparametric tests gain relatively more power outperforming the T^2 test with higher sample sizes. With higher p:s, the powers of the nonparametric tests also increase, but the powers do not exceed the power of Hotelling's T^2 test. The relative order of the power curves of the nonparametric tests was not found to be significantly affected by an increase in sample size. As for the number of dimensions, varying the dimension count was not found to have a significant impact on powers in general, but with lower p:s, Hotelling's T^2 test gains some power with increasing d, whereas the marginal sign tests Q^2_{MS} and Q^2_{ICMS} and the signed-rank test Q^2_{HPTR} based on Tyler's angles lose power. With higher p:s, the power curves remained largely unaffected by a change in d.

Based on these simulations, it can be concluded that – in a practical situation – the choice of the location test matters and the choice should largely depend on the distribution of the data. In practice, one does not know the underlying distribution. An important first step then in choosing which test to use would be to look at a graphical representation of the data at hand (scatterplots, for example). A look at the figures in Chapter 3 might help in deciding which test to use. The closer the distribution is to an elliptic distribution ($p = 2$), the more recommendable it is to use Hotelling's T^2 test. However, Hotelling's T^2 test loses power quickly as the distribution deviates from ellipticity. In those cases, the use of the sign tests is recommended. The Oja sign test Q^2_{OS} displays highest powers under non-elliptic distributions. However, the information on powers regarding Oja tests is limited due to computational

issues. The spatial sign test Q^2_{SS} using inner standardization and the sign test Q^2_{HPTS} based on Tyler's angles come close in power. Using the marginal sign tests Q^2_{MS} and Q^2_{ICMS} is generally not recommended since they yield almost uniformly lower power than the other sign tests.

If Hotelling's T^2 test is valid (the first two moments exist), then the use of the signed-rank tests is generally not recommended since those tests do best with distributions under which Hotelling's T^2 test is the optimal choice and they get almost uniformly outperformed by the sign tests under non-elliptic densities. The signed-rank tests do, however, come close to Hotelling's T^2 test in power under elliptic densities so if Hotelling's T^2 test is not valid, then the use of the signed-rank tests is recommended over the sign tests. The spatial signed-rank test Q^2_{SR} using inner standardization and the signed-rank test Q^2_{HPTR} based on Tylers's angles generally perform best. Also, the power curves of the signed-rank tests converge as p increases so one should not worry too much about deciding which signed-rank test to use.

The sign and signed rank tests Q^2_{HPTS} and Q^2_{HPTR} based on Tyler's angles behaved accordingly with the other sign and signed-rank whether or not the underlying distribution followed the elliptic model. Thus, based on these simulations, if one chooses to use these tests, one should not worry too much about whether the ellipticity requirement is met. Similarly, the marginal sign and signed-rank tests Q^2_{ICMS} and Q^2_{ICMR} in the symmetric IC model behaved accordingly with the other sign and signed-rank tests whether or not the underlying distribution followed the symmetric IC model.

In general, the higher the sample size, the more recommendable it is to use the nonparametric tests. Based on these simulations, sample size does not affect which nonparametric test to use, because the relative order of the power curves does not change with increasing sample size. The results also show that the dimension of the data does not play a significant role in determining which location test to choose. Increasing dimension improves the relative power of Hotelling's T^2 test under non-elliptic densities and decreases the powers of the marginal signs tests Q^2_{MS} and Q^2_{ICMS} and those of the spatial signed-rank test Q^2_{HPTR} based on Tyler's angles, but these results do not affect the suggestions given.

Overall, the study does not provide a clear-cut answer as to what is the best test, because no test exhibits highest powers uniformly. In other words, there is no "free lunch". On the basis of these simulations results as a whole, the use of the spatial sign test Q^2_{SS} using inner standardization or the sign test Q^2_{HPTS} based on Tyler's angles is suggested under non-elliptic densities. The performance of the Oja tests remains partially unclear due to computational issues. Under elliptic densities, the use of Hotelling's T^2 test is recommended. However, Hotelling's T^2 test exhibits bias with smaller sample sizes as does the spatial signed-rank test Q^2_{SR} using inner standardization. If Hotelling's T^2 test is not valid, then the use of the signed-rank test Q^2_{HPTR} based on Tyler's angles is recommended. With higher sample sizes, the powers of the nonparametric tests come close to the power of the T^2 test under elliptic densities.

One strength of this thesis is the extensive simulation study. The performance of the Oja tests needs further investigation. Future studies might consist of expanding the simulation study by using other distributions than L_p-norm distributions. Also, adding outliers to data could be used to study the robustness of the tests. In addition, permutation testing might be implemented.

References

[1] Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Hobocken, US, 3rd edition.

[2] Chakraborty, B. and Chaudhuri, P. (1996). On a Transformation and Retransformation Technique for Constructing Affine Equivariant Multivariate Median. *Proceedings of the American Mathematical Society*, 124, 2539−2547.

[3] Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating Transformation Retransformation on the Spatial Median and Angle Test. *Statistica Sinica*, 8, 767−784.

[4] Fischer, D., Möttönen, J., Nordhausen, K., and Vogel, H. (2013). OjaNP: Multivariate Methods Based on the Oja Median and Related Concepts. R package vesion 0.9-6. URL http://cran.r-project.org/web/packages/OjaNP/index.html.

[5] Gentle, J. (2003). *Random Number Generation and Monte Carlo Methods*. Springer.

[6] Gupta, A.K. and Song, D. (1997). L_p-norm Spherical Distribution. *Journal of Statistical Planning and Inference*, 60, 241−260.

[7] Hallin, M. and Paindaveine D. (2002a). Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks. *The Annals of Statistics*, 30, 1103−1133.

[8] Hallin, M. and Paindaveine D. (2002b). Randles' Interdirections or Tyler's Angles? In Y. Dodge, Ed., *Statistical data analysis based on the L_1-norm and related methods*, 271−282.

[9] Hettmansperger, T.P., Nyblom, J., and Oja, H. (1994). Affine Invariant Multivariate One-sample Sign Tests. *Journal of the Royal Statistical Society*, 56, 221−234.

[10] Hettmansperger, T.P., Möttönen, J., and Oja, H. (1997). Affine Invariant Multivariate One-sample Signed-rank Tests. *Journal of the American Statistical Association*, 92, 1591−1600.

[11] Hettmansperger, T.P. and Randles, R.H. (2002). A Practical Affine Equivariant Multivariate Median. *Biometrica*, 89, 851−860.

[12] Hotelling, H. (1931). The Generalization of Student's Ratio. *Annals of Mathematical Statistics*, 2(3), 360−378.

[13] Liang, J. and Ng, K.W. (2008). A Method for Generating Uniformly Scattered Points on the L_p-norm Unit Sphere and Its Applications. *Metrika*, 68, 83−98.

[14] Magyar, A. and Tyler, D. (2011). The Asymptotic Efficiency of the Spatial Median for Elliptically Symmetric Distributions. *Sankhya*, Series B, 73, 188−191.

[15] Möttönen, J. and Oja, H. (1995). Multivariate Spatial Sign and Rank Methods. *Journal of Nonparametric Statistics*, 5, 201−213.

[16] von Neumann, J. (1951). Various Techniques Used in Connection With Random Digits. Monte Carlo methods. *National Bureau of Standards Applied Math Series*, 12, 36−38.

[17] Nordhausen, K., Oja, H., and Tyler, D.E. (2006). *On the Efficiency of Invariant Multivariate Sign and Rank Tests*. In Liski, E.P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G.P.H. (editors) "Festschrift for Tarmo Pukkila on his 60th birthday", 217−231, University of Tampere, Tampere.

[18] Nordhausen, K., Oja, H., and Tyler, D.E. (2008). Tools for Exploring Multivariate Data: The Package ICS. *Journal of Statistical Software*, 28(6), 1−31.

[19] Nordhausen, K., Oja H., and Paindaveine, D. (2009). Signed-ranks Tests for Location in the Symmetric Independent Component Model. *Journal of Multivariate Analysis*, 100(5), 821−834.

[20] Nordhausen, K., Möttönen, J., and Oja, H. (2011). MNM: Multivariate Nonparametric Methods. An Approach Based on Spatial Signs and Ranks. R package vesion 1.0−0. URL http://cran.r-project.org/web/packages/MNM/index.html.

[21] Nordhausen, K., Sirkiä, S., Oja, H., and Tyler, D.E. (2012). ICSNP: Tools for Multivariate Nonparametrics. R package vesion 1.0−9. URL http://cran.r-project.org/web/packages/ICSNP/index.html.

[22] Nordhausen, K., Oja, H., and Tyler, D.E. (2013). ICS: Tools for Exploring Multivariate Data via ICS/ICA. R package vesion 1.2−4. URL http://cran.r-project.org/web/packages/ICS/index.html.

[23] Oja, H. (1983). Descriptive Statistics for Multivariate Distributions. *Statistics & Probability Letters*, 1, 327−332.

[24] Oja, H. (1999). Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: A review. *Scandinavian Journal of Statistics*, 26, 319−343.

[25] Oja, H. (2010). *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks*. Springer.

[26] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33, 1065.

[27] Puri, M.L. and Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York, USA.

[28] Randles, R. H. (1989). A Distribution-free Multivariate Sign Test on Interdirections. *Journal of the American Statistical Association*, 84, 1045−1050.

[29] Ronkainen, T., Oja, H., and Orponen, P. (2003). Computation of the Multivariate Oja Median. *Developments in Robust Statistics*, 344−359.

[30] Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27, 832.

[31] Serfling, R.J. (2006). Multivariate Symmetry and Asymmetry. *Encyclopedia of Statistical Sciences, Second Edition* (S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic, eds.), 8, 5338−5345.

[32] Sinz F. H., Gerwinn, S., and Bethge M. (2009). Characterization of the p-Generalized Normal Distribution. *Journal of Multivariate Analysis*, 100(5), 817−820

[33] Song, D. and Gupta, A.K. (1997). L_p-norm Uniform Distribution. *Proceedings of the American Mathematical Society*, 125, 595−601.

[34] Tyler, D.E. (1987). A distribution-free M-estimator of Multivariate Scatter. *Annals of Statistics*, 15, 595−601.

[35] Visuri, S., Koivunen, V. and Oja, H. (2000). Sign and Rank Covariance Matrices. *Journal of Statistical Planning and Inference*, 91, 557−575.

[36] Zuo, Y. (1998). Contributions to the Theory and Applications of Statistical Depth Functions. Ph.D. dissertation, University of Texas, Dallas.

Appendix: Simulation plots

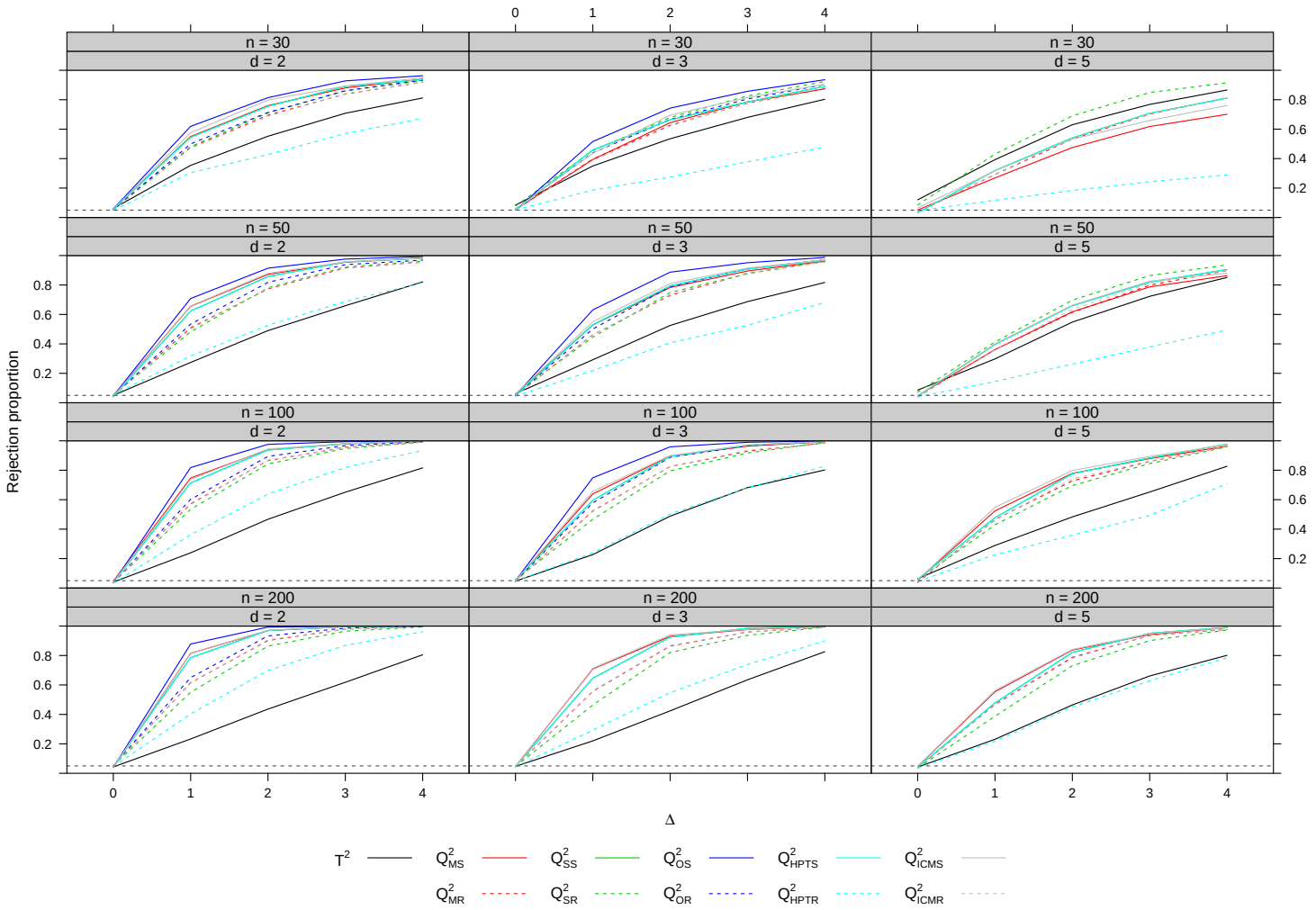

Figure 1. Rejection proportions (for $n = 30$, 50, 100, 200 and $d = 2$, 3, 5, based on 1,000 replications) of all tests for data coming from the p-generalized normal distribution, $p = 0.5$.

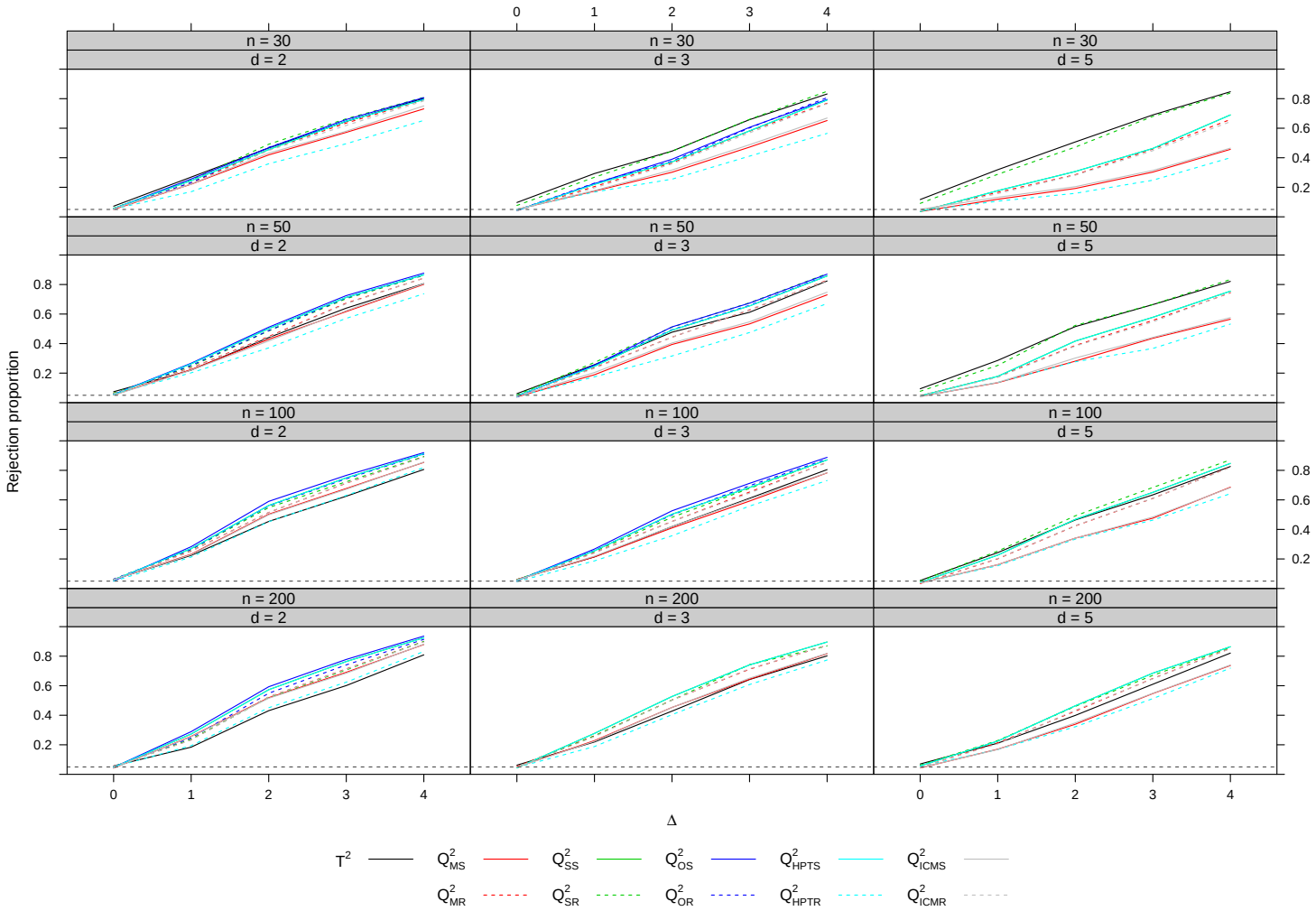

Figure 2. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the p-generalized normal distribution, $p = 1$.

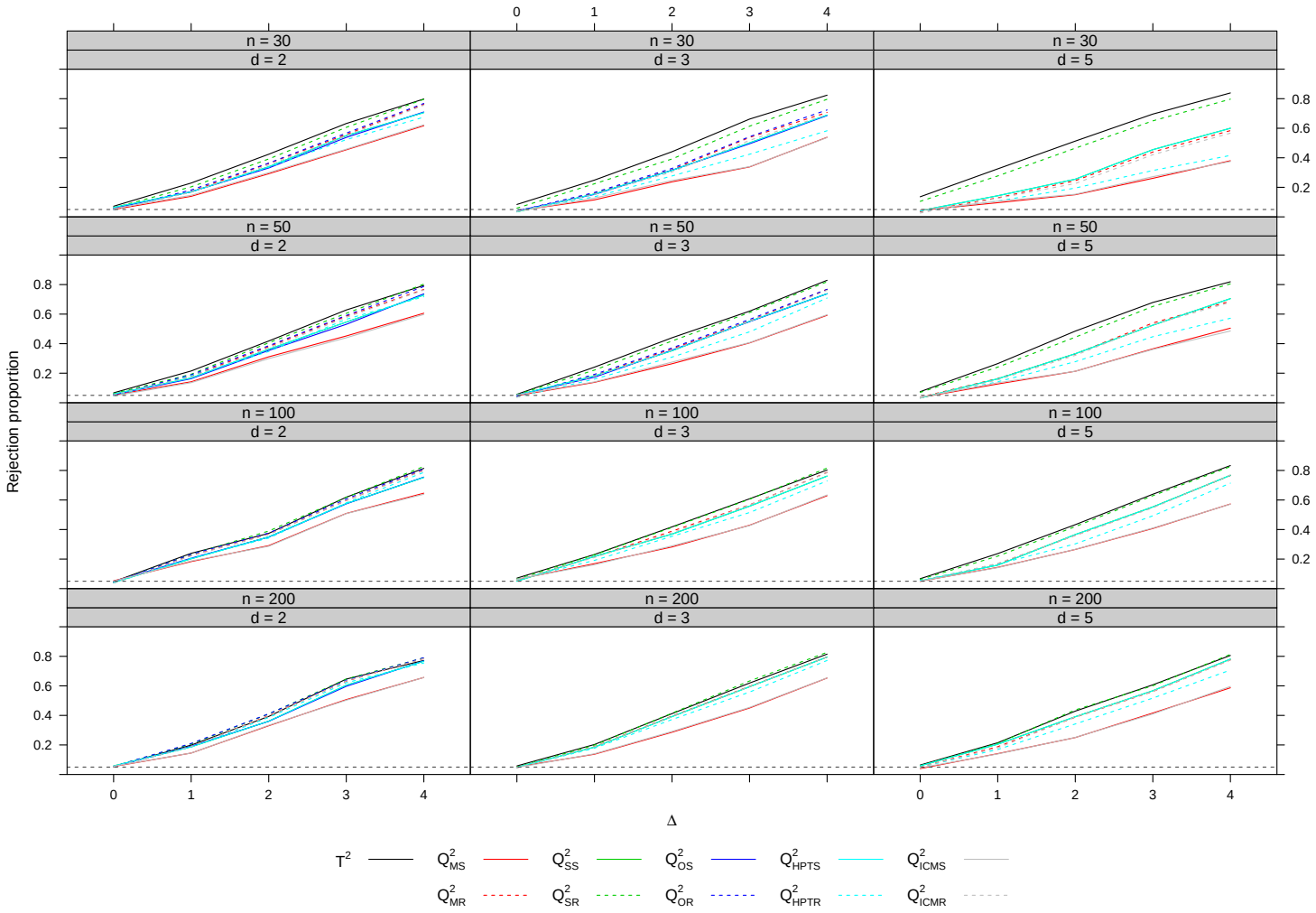

Figure 3. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the p-generalized normal distribution, $p = 1.5$.

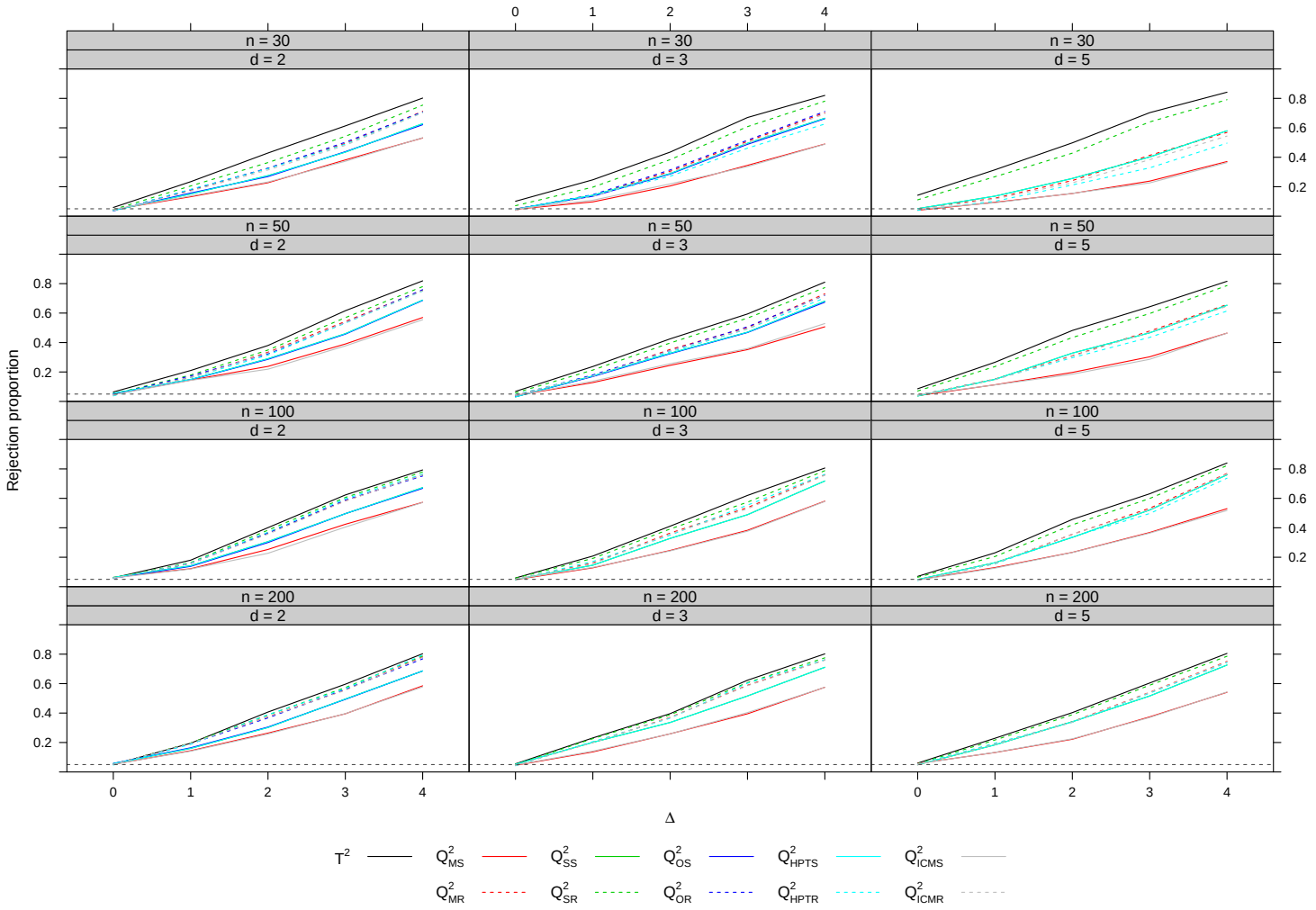

Figure 4. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the p-generalized normal distribution, $p = 2$.

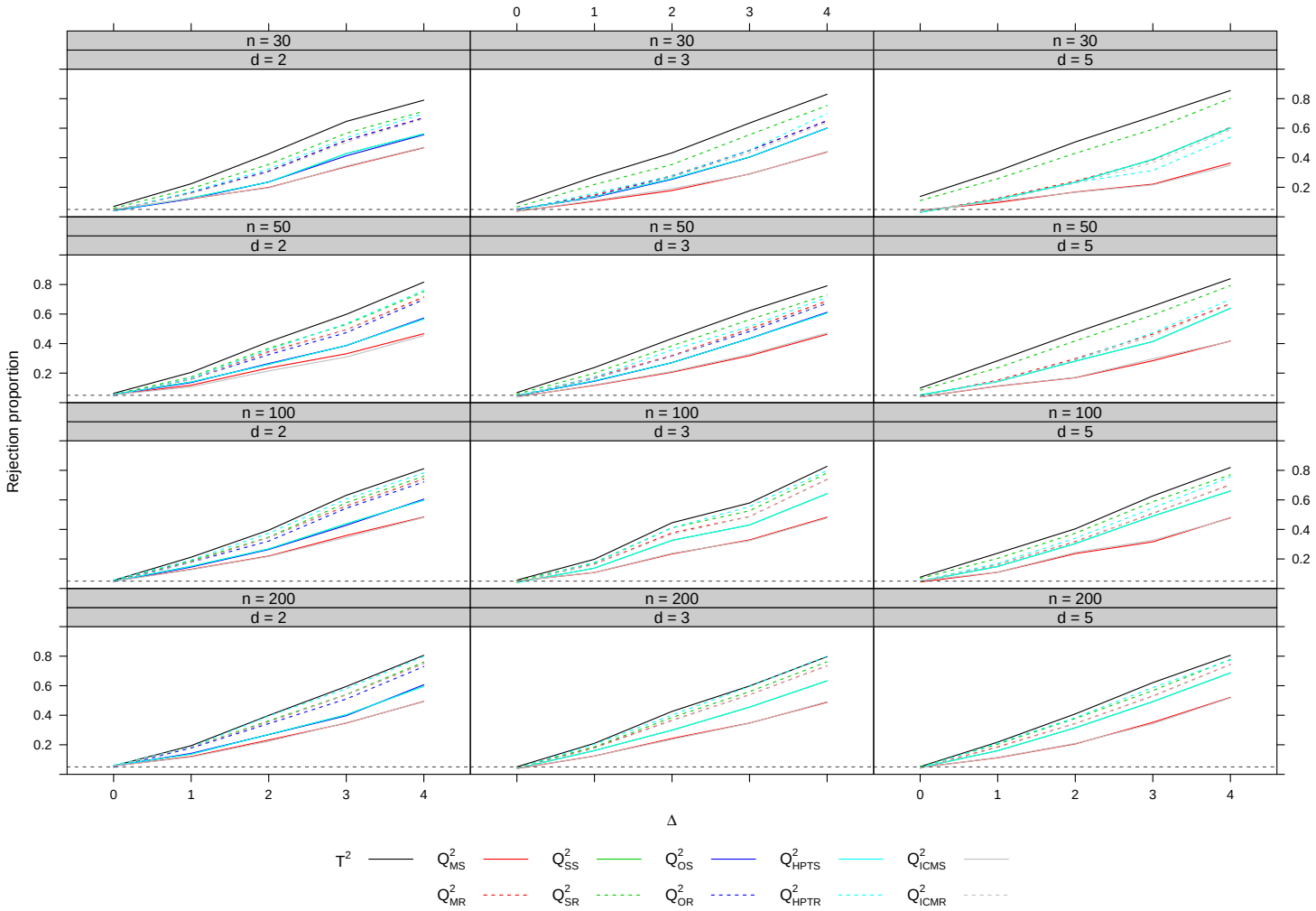

Figure 5. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the p-generalized normal distribution, $p = 3$.

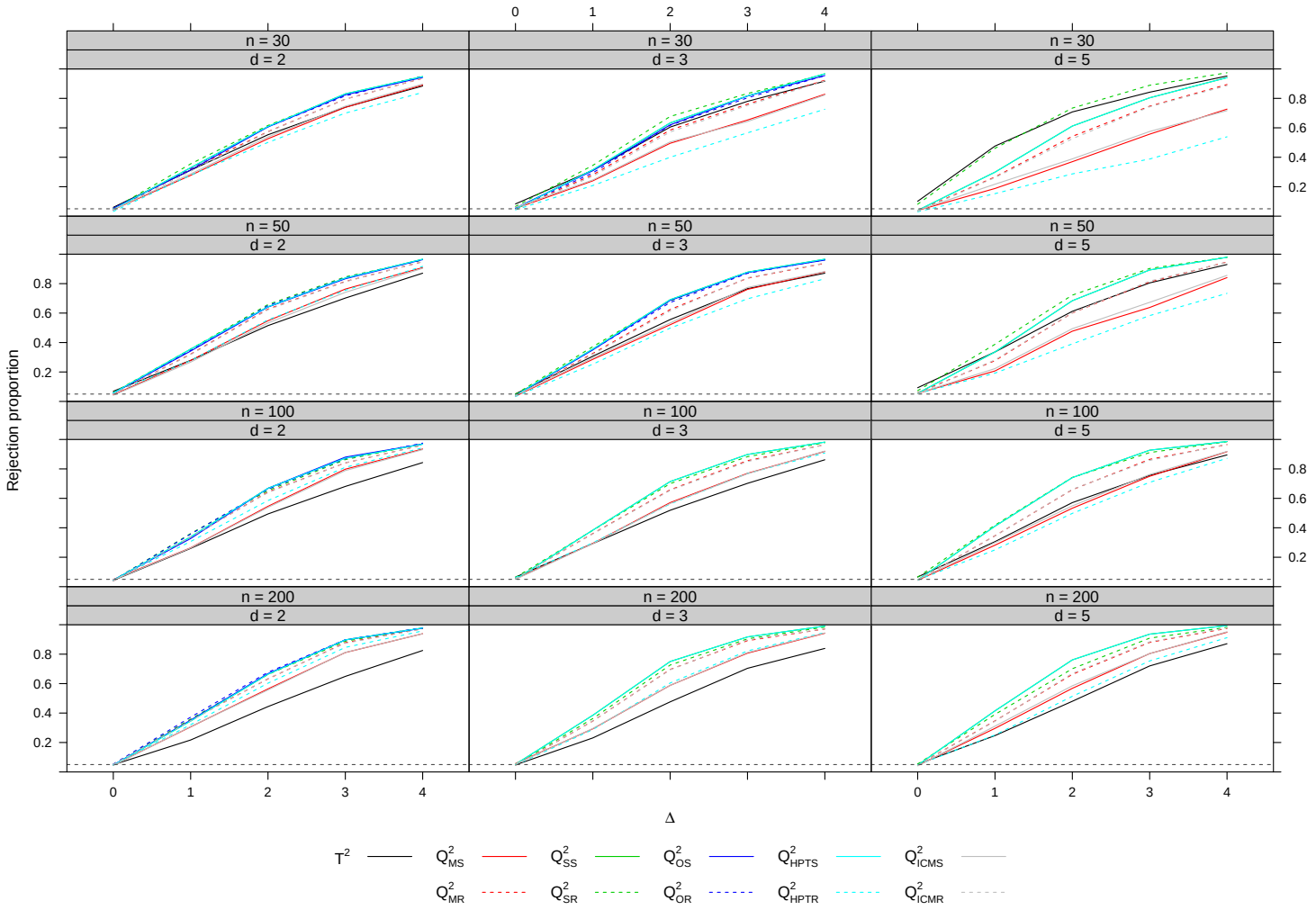

Figure 6. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 2$ with $df = 3$.

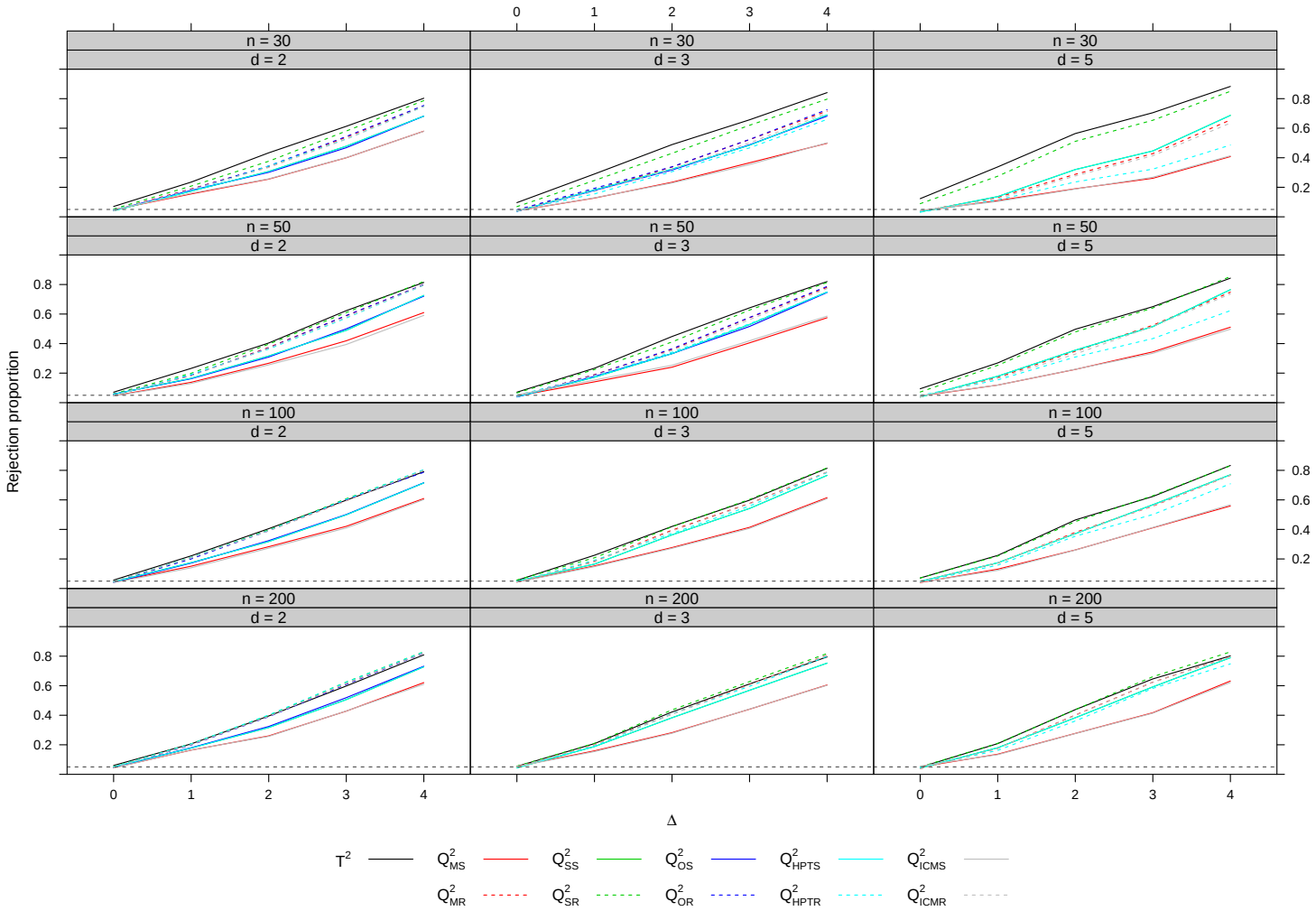

Figure 7. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 3$ with $df = 3$.

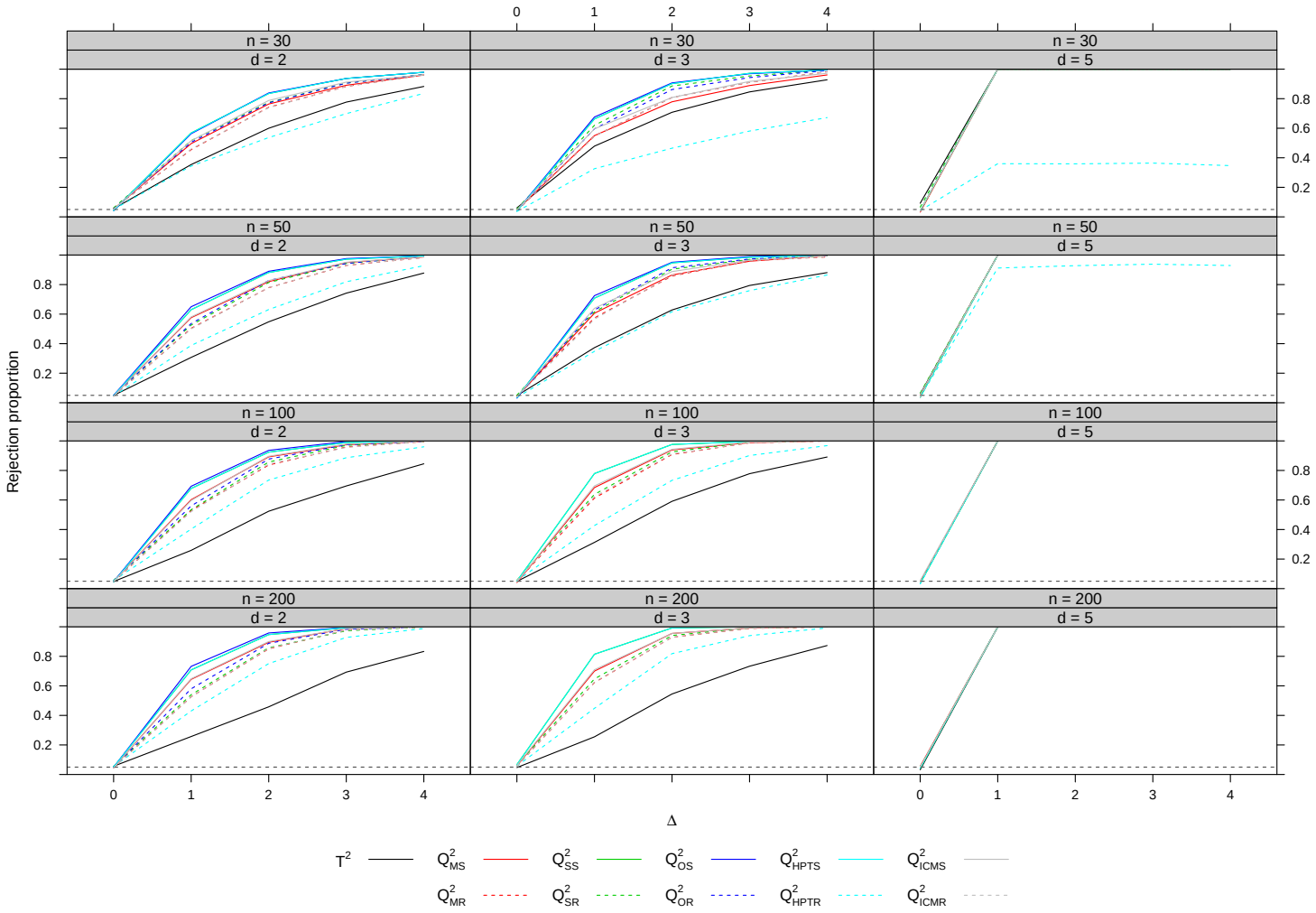

Figure 8. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 1$ with $df = 9$.

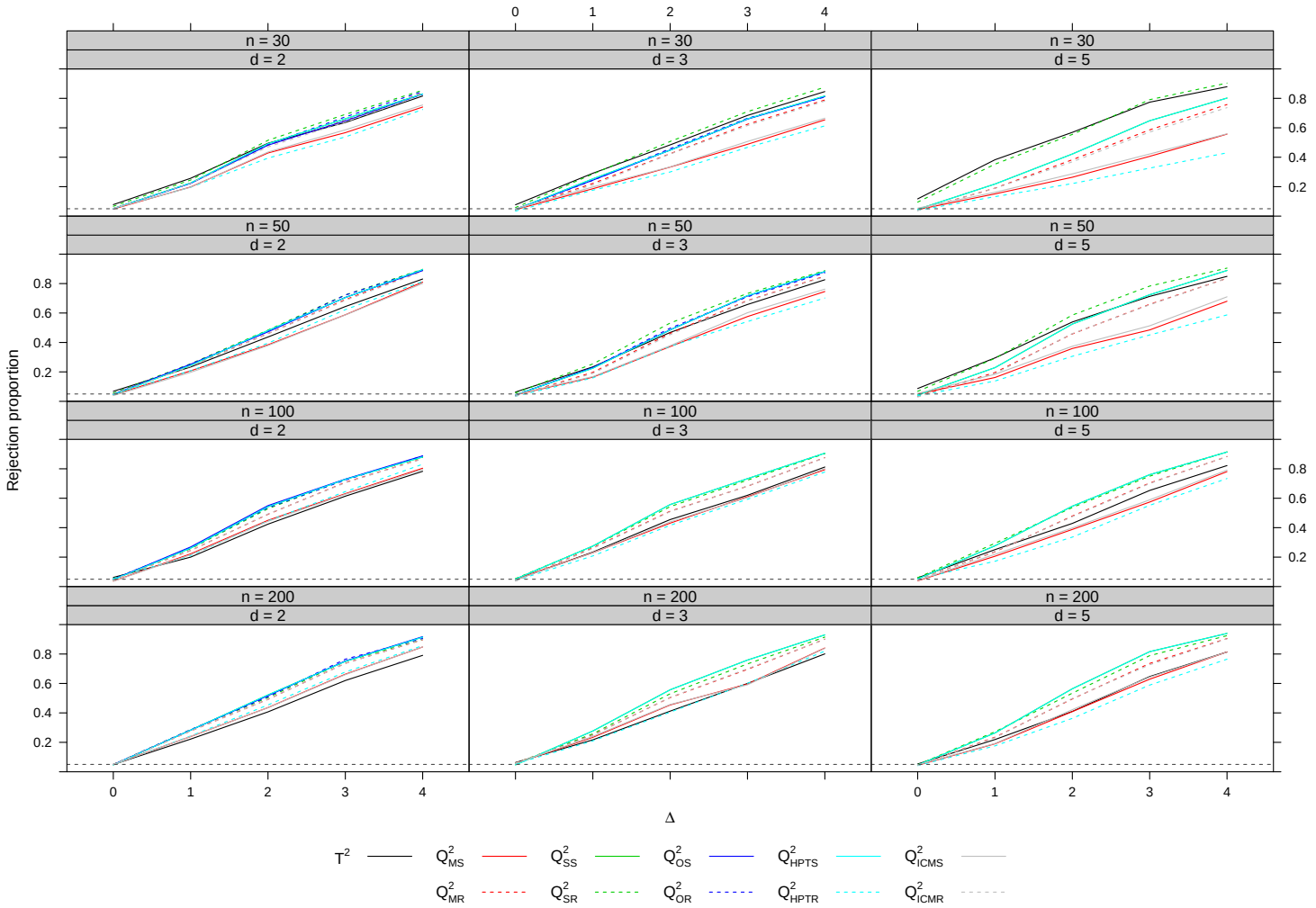

Figure 9. Rejection proportions (for $n = 30, 50, 100, 200$ and $d = 2, 3, 5$, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 1.5$ with $df = 9$.

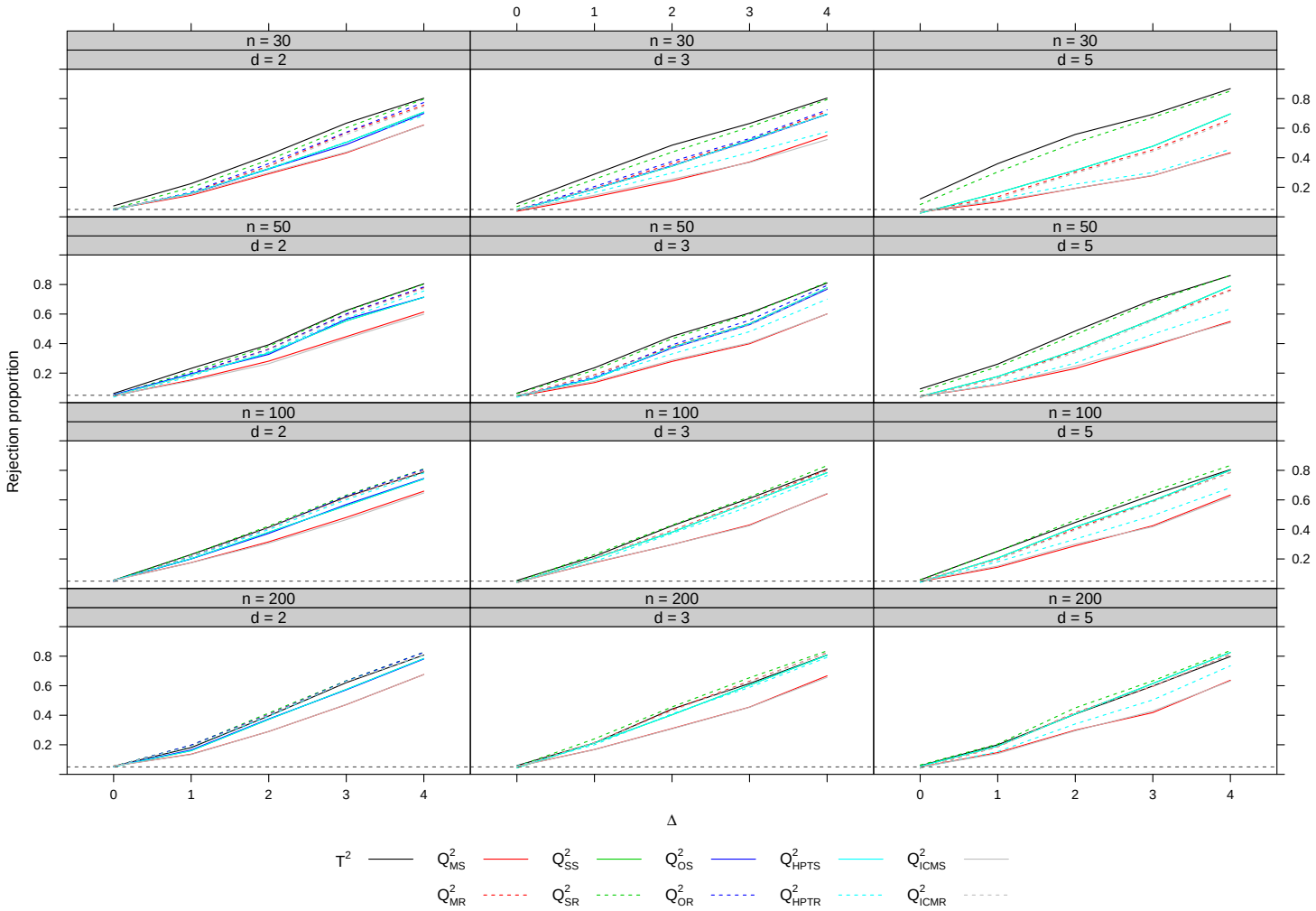

Figure 10. Rejection proportions (for $n = 30$, 50, 100, 200 and $d = 2$, 3, 5, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 2$ with $df = 9$.

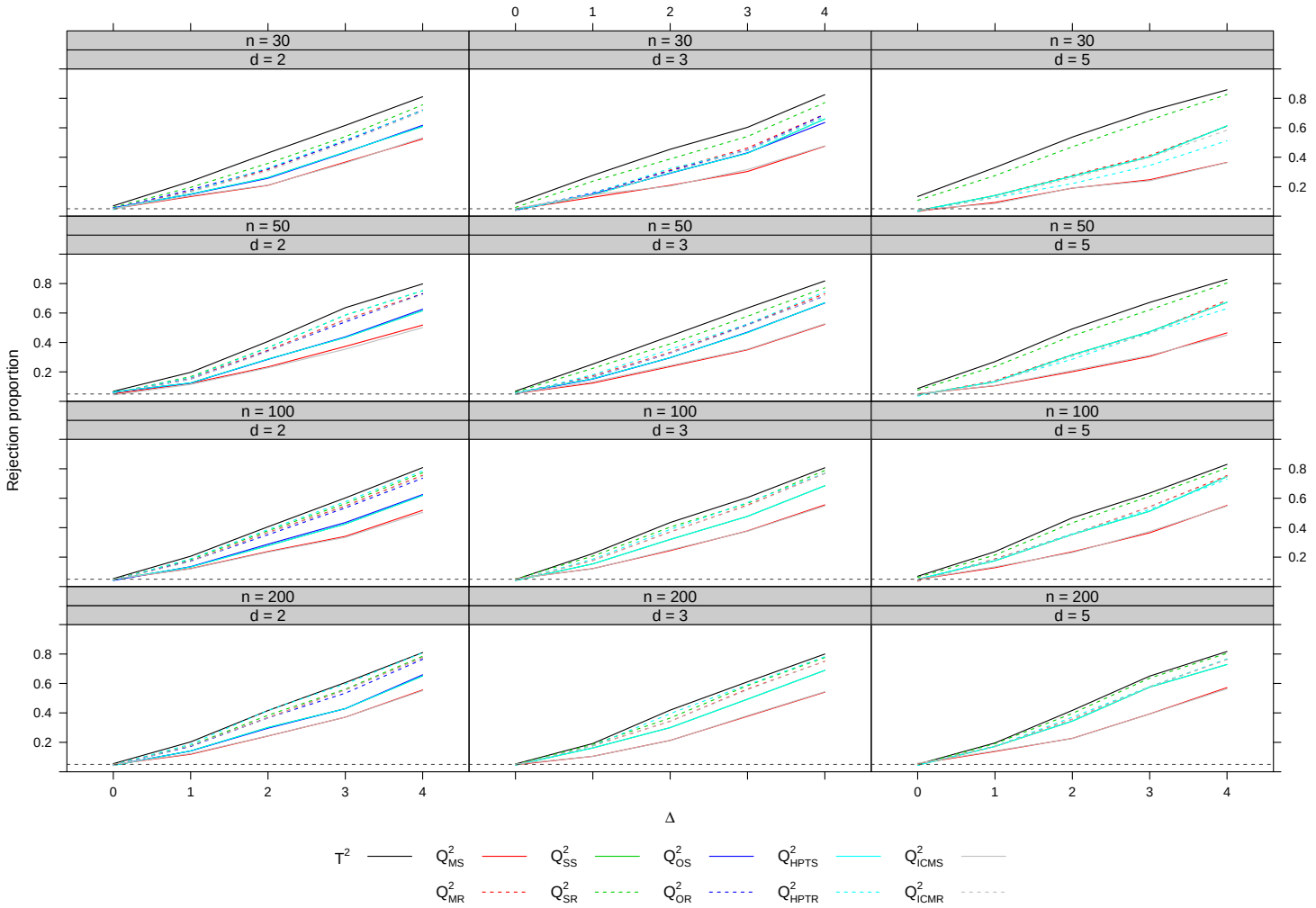

Figure 11. Rejection proportions (for $n = 30$, 50, 100, 200 and $d = 2$, 3, 5, based on 1,000 replications) of all tests for data coming from the L_p-norm multivariate t-distribution, $p = 3$ with $df = 9$.